AERA Open
October-December 2020, Vol. 6, No. 4, pp. 1–18
DOI: 10.1177/2332858420977208
Article reuse guidelines: sagepub.com/journals-permissions

The Bias-Variance Tradeoff: How Data Science Can Inform Educational Debates

Shayan Doroudi
University of California, Irvine

In addition to providing a set of techniques to analyze educational data, I claim that data science as a field can provide broader insights to education research. In particular, I show how the bias-variance tradeoff from machine learning can be formally generalized to be applicable to several prominent educational debates, including debates around learning theories (cognitivist vs. situativist and constructivist theories) and pedagogy (direct instruction vs. discovery learning). I then look to see how various data science techniques that have been proposed to navigate the bias-variance tradeoff can yield insights for productively navigating these educational debates going forward.

Keywords: *bias-variance tradeoff, learning theories, direct instruction, discovery learning, research methods, epistemology, artificial intelligence, machine learning*

Introduction

In recent years, data science and machine learning have been used to make sense of various forms of educational data. Educational data science can enhance our understanding of learning and teaching; give insights to teachers, students, and administrators about learning happening in the classroom; and lead to the creation of data-driven adaptive learning systems. However, I claim that machine learning and data science have more to offer than a set of techniques that can be applied to educational data. Theoretical concepts and principles in machine learning can provide broader insights to education research. In particular, in this article, I show that the bias-variance tradeoff from machine learning can provide a new lens with which to view prominent educational debates.

The field of education is filled with seemingly perennial debates that have important implications on the nature of education research and practice. These debates span epistemology (e.g., positivism vs. constructivism), ontology (e.g., cognitive vs. situative theories of learning), methodology (e.g., quantitative vs. qualitative), and pedagogy (e.g., direct instruction vs. discovery learning). Interestingly, many researchers have contended that these debates are often premised on false dichotomies and admit that a moderate position is more warranted (Alfieri et al., 2011; A. L. Brown & Campione, 1994; Cobb, 2007; Doroudi et al., 2020; Greeno, 2015; Johnson & Onwuegbuzie, 2004; Kersh & Wittrock, 1962; Shaffer, 2017; Shaffer & Serlin, 2004; Tobias & Duffy, 2009), yet these debates persist and many researchers still seem to favor one side over the other, resulting in somewhat isolated research communities that speak past each other (if they speak to each other at all; Jacobson et al., 2016).

In this article, I show that the bias-variance tradeoff can be formally generalized to help explain the nature of these debates in the education literature. I first introduce the bias-variance tradeoff and how it is used in statistics and machine learning. I then describe a formal generalization of the bias-variance decomposition and map this generalized version to debates around theories of learning, methodology, and pedagogy, while also showing how these tradeoffs are rooted in similar debates in the history of artificial intelligence (AI). While the first two sections that introduce and formalize the bias-variance tradeoff are somewhat technical, later sections of the article primarily make descriptive claims and use representative quotations from prominent researchers in these debates. As such, the article aims to be accessible to and of interest to both quantitative and qualitative education researchers. We will then look towards how data scientists and machine learning researchers have productively navigated the bias-variance tradeoff, in order to find several insights on how to navigate the corresponding educational debates. I finally discuss how the bias-variance tradeoff can help us understand overarching themes across these debates, including how debates around epistemology in education research could explain why these debates are ongoing, despite pragmatic attempts to overcome them.

By mapping the bias-variance tradeoff to these educational debates, I hope to accomplish several goals. First, it can help us understand the relationships between various educational debates (e.g., on learning theories, methodology, pedagogy, and epistemology) as well as debates in other

fields (e.g., machine learning, AI, and linguistics). Second, it provides a novel way to formally rationalize the comparative advantages of both sides in each debate, rather than using strawman arguments. Third, it can help us understand surprising connections between different approaches that may seem unrelated at first glance. For example, it can help us explain the connection between situativist and constructivist theories, as well as the connection between these theories and neural networks. On a more meta level, this articel shows how a quantitative technique (the bias-variance tradeoff) can give insights on qualitative methods. Finally, related to the previous point, it can help us see connections between pragmatic techniques for navigating the bias-variance tradeoff in data science and analogous techniques for productively navigating these educational debates. In short, the purpose of this article is not to resolve these perennial debates in education but rather to provide a framework with which we can better understand them and hopefully carry more meaningful conversations going forward.

The Bias-Variance Tradeoff

The bias-variance tradeoff was first formally introduced by Geman et al. (1992). It refers to the fact that when trying to make a statistical prediction (e.g., estimating a parameter of a distribution or fitting a function), there is a tradeoff between the accuracy of the prediction and its precision, or equivalently between its bias (opposite of accuracy) and variance (opposite of precision). Many education researchers may be familiar with the concepts of bias and variance in terms of validity and reliability in measurement. In this section, I will first informally explain the notion of bias and variance using targets, and then explain the notion more precisely as it has been used in the machine learning literature. This section and the one that follows provide the necessary technical background in a way that is meant to be accessible to education researchers but not explicitly tied to concerns in education research. However, these sections make way for a more informed discussion of how these ideas extend to education.

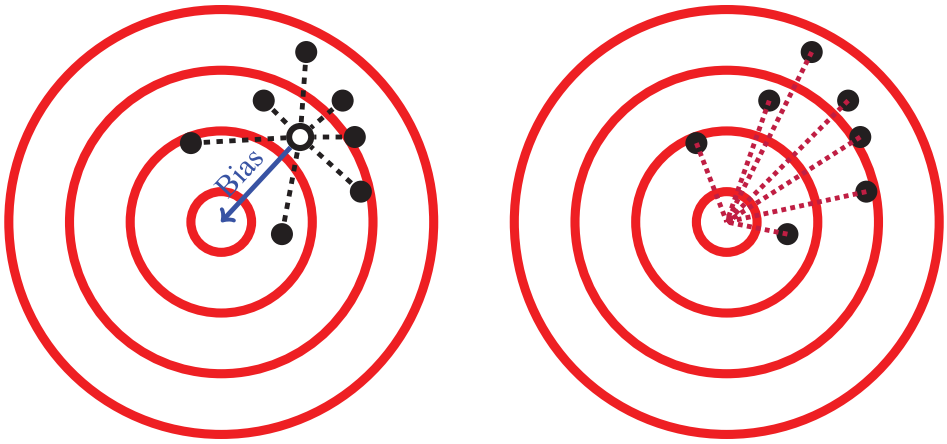
Suppose a process randomly generates points on a target (e.g., an archer shooting arrows at the target) as shown in Figure 1a. The bias of the process is how far the points are from the center of the target on average, as depicted by the blue solid line, and the variance of the process is a measure of how far the points are to the centroid of the points on average, as depicted by the black dotted lines. Precisely, the variance can be estimated by taking the mean of the squared distances from each point to the centroid of the points (the dotted lines shown in Figure 1a). Figure 2 shows examples of low bias and low variance, high bias and low variance, low bias and high variance, and both high bias and high variance. Now suppose an archer has only one shot at the target and wants to be as close to the center as possible. One way to measure how far the archer's shot will likely be from the center is the mean squared error, which is the average distance from a point to the center of the target, which can be estimated as the mean of the squared distances between all the points and the center of the target as shown in Figure 1b. The mean squared error increases as the bias or variance increases. In fact, a well-known result is that the mean squared error decomposes into a bias and variance term:

$$\text{Mean Squared Error} = \text{Bias}^2 + \text{Variance}.$$

In the context of machine learning, the goal is to estimate a function that minimizes the mean squared error distance between the estimated function and the true function. To fit a function, an algorithm searches for a best fitting function (or *estimator*) in a *function class*. A function class is a set of functions that usually have the same underlying form but must be instantiated with particular parameters. For example, the class of linear estimators using a particular set of features (fit via linear regression) is one function class. The bias of a function class represents how different the best fitting model in the function class is from the target function. For example, if we wanted to predict y where $y = 3.5x^2 - x$ using a linear estimator, we could never fit the curve perfectly, and thus, the function class would have some bias. On the other hand, if we were to use a function class of quadratic estimators (i.e., all functions of the form $y = ax^2 + bx + c$), then we could fit the target perfectly, and so the function class would have no bias. Variance represents the amount of variability in functions of the function class, or in other words is a measure of the complexity or size of the function class. The function class of quadratic estimators has higher variance than linear estimators because quadratic functions are capable of modeling more complex patterns; in addition to being able to model all lines, they can also approximate "U"-shaped functions. The function class of all polynomials has much higher variance. Therefore, using a function class of all polynomials to fit $y = 3.5x^2 - x$ will naturally have higher mean squared error than using the function class of quadratic estimators, even though both have zero bias.

Thus, to effectively use machine learning one tries to use a function class that balances the bias-variance tradeoff. Ideally, we would use the smallest function class that can capture the target function. However, given that we do not know the form of the target function ahead of time, this is not always possible.[1]

We will now formally express the notion of bias and variance. Suppose we are trying to predict some target function $f : \mathbb{R}^m \to \mathbb{R}$. Let $y = f(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_m)$. We use a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{P}_D$ to fit an estimator $\hat{f}$ of the target *f* using some statistical algorithm, which (possibly implicitly) searches over some function class $\mathcal{F}$ to find the best $\hat{f}$ in the function class. For any $\mathbf{x} \in \mathbb{R}^m$, the bias of the predictor's function class is the difference between the expected value of the predictor at $\mathbf{x}$ and the expected value of the target at $\mathbf{x}$ (which we will assume is not random):



(a) Bias and variance. (b) Mean-squared error.

FIGURE 1. *A depiction of the bias-variance tradeoff using targets. If these points are thought of as arrows, then the goal would be for the points to be near the center of the target. (a) To the extent which the points are far from the center, they suffer from bias (solid blue line) and/or variance (dashed black lines). (b) The bias and variance combine to form mean squared error (dotted purple lines).*

$$\mathbb{E}_{\mathcal{P}_D}[\hat{f}(\mathbf{x})] - f(\mathbf{x}).$$

The variance of the predictor's function class (for a given $\mathbf{x}$) is the expected difference between the value of the predictor estimated on a randomly sampled data set and the expected value of the predictor,

$$\mathbb{E}_{\mathcal{P}_D}[(\hat{f}(\mathbf{x}) - \mathbb{E}_{\mathcal{P}_D}[\hat{f}(\mathbf{x})])^2].$$

A standard measure for the error in a prediction is the mean squared error,

$$\mathbb{E}_{\mathcal{P}_D}[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2].$$

We can now formally express the bias-variance decomposition of the mean squared error[2]:

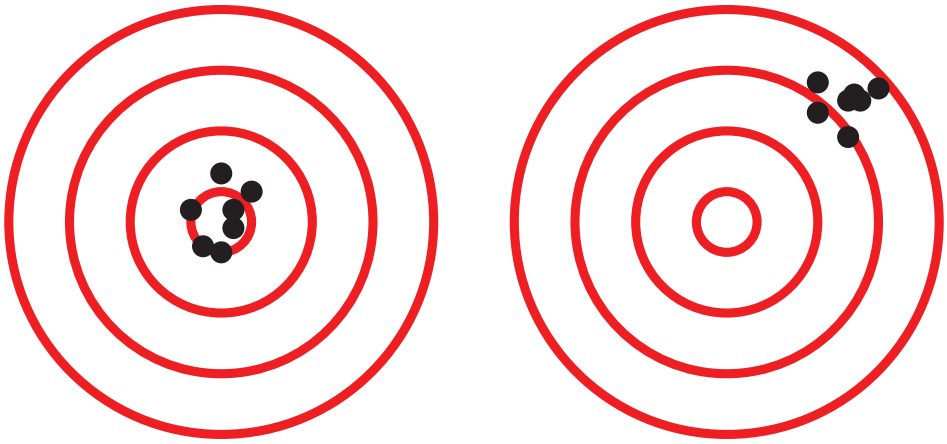
Theorem 1 (Bias-Variance Decomposition). *Under the setting described above, the mean squared error decomposes into a bias term and variance term as follows*:

$$\underbrace{\mathbb{E}_{\mathcal{P}_D}[(\hat{f}(\mathbf{x} - f(\mathbf{x}))^2]}_{\text{Mean Squared Error}} = \underbrace{(\mathbb{E}_{\mathcal{P}_D}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2}_{\text{Bias Squared}} + \underbrace{\mathbb{E}_{\mathcal{P}_D}[(\hat{f}(\mathbf{x}) - \mathbb{E}_{\mathcal{P}_D}[\hat{f}(\mathbf{x})])^2]}_{\text{Variance}}.$$

The bias-variance decomposition gives us a sense of why there is a bias-variance tradeoff. For two function classes that have the same prediction error, if one function class is more biased than the other, then we know the other must have higher variance. Naturally, many good machine learning techniques tend to be more susceptible to either bias or variance. Figure 3 shows how bias, variance, and mean squared error tend to change as a a function of model complexity.[3] Increasing model complexity could correspond to increasing the number of features in a linear regression, increasing the highest degree in a polynomial regression, or increasing the number of layers in a neural network. As the complexity increases, the bias tends to decrease but the variance tends to increase. There is typically a degree of complexity where the mean squared error is minimized by effectively balancing bias and variance. Many techniques have been proposed to navigate this bias-variance tradeoff. We will discuss some of these techniques at the end of this article, in the context of navigating tradeoffs that exist in education.

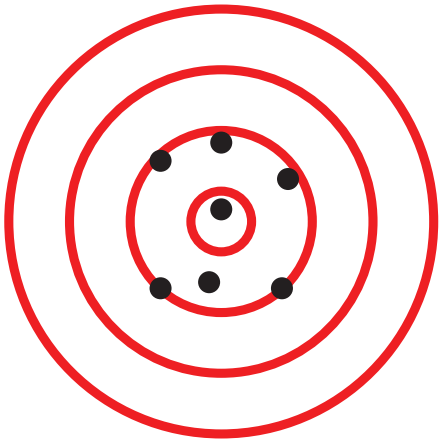
The Generalized Bias-Variance Decomposition

Notice that in the target diagrams, the source of bias and variance is due to the random mechanism by which the points were generated. This randomness could be due to random fluctuations in a (novice) archer's aim. In the case of machine learning, this randomness comes from the data used to estimate the function. Notice that in the archer's case there are no data present. While in the machine learning literature, the bias-variance tradeoff is often represented as being about overfitting or underfitting to data, a key insight of this article is that *the bias-variance tradeoff is not really about data but rather a property of any random mechanism that tries to approximate some target*. This can be shown more precisely by noting that the proof of the bias-variance decomposition is agnostic to the source of randomness; the decomposition holds regardless of the probability distribution the expectations are taken over. In the case of machine learning, this may be due to a data set used to fit a function, but in other cases, like that of an archer or that of a teacher (as discussed in the next section), the randomness could be due to an archer's aim or the chaotic stochasticity of a typical classroom environment,



(a) Low bias, low variance.

(b) High bias, low variance.



(c) Low bias, high variance.

(d) High bias, high variance.

FIGURE 2. *A depiction of varying combinations of bias and variance: (a) low bias and low variance, (b) high bias and low variance, (c) low bias and high variance, and (d) high bias and high variance.*

respectively. This leads us to a generalization of the bias-variance decomposition:

Theorem 2 (Generalized Bias-Variance Decomposition). *Suppose our goal is to approximate some target:* $T : \mathbb{R}^m \to \mathbb{R}^n$. *Let* $\mathcal{M}$ *be a stochastic mechanism (i.e., a partially random process) that randomly chooses a function* $\hat{T} : \mathbb{R}^m \to \mathbb{R}^n$ *from a function class* $\mathcal{T}$ *(i.e.,* $\hat{T} \sim \mathcal{M}$*). Then we have the following for all* $\mathbf{x} \in \mathbb{R}^m$*:*

$$\underbrace{\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - T(\mathbf{x}))^2]}_{\text{Mean Squared Error}} = \underbrace{(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))^2}_{\text{Bias Squared}} + \underbrace{\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2]}_{\text{Variance}}.$$

A proof of the theorem is given in Appendix A. The only real difference between Theorem 1 and Theorem 2 is that the stochasticity of the mechanism $\mathcal{M}$ is not necessarily due to probability distributions over data sets, that is, $\hat{T}$ need not be learned from data. Table 1 shows how the various variables in this generalized bias-variance tradeoff (Theorem 2) map onto the specific bias-variance tradeoff in machine learning (Theorem 1).[4]

The Bias-Variance Tradeoff in Education

We can now analyze how the bias-variance tradeoff applies to many educational debates. In what follows, we will discuss several prominent educational debates, focusing primarily on debates around learning theory and pedagogy (but also bringing up connections to other related debates). In each section, I first give a summary of the debates and then examine how the generalized bias-variance decomposition can help characterize the different positions in the debates. I rely heavily on historical examples from the history of AI and the learning sciences and give illustrative quotes from researchers on either side of the debates. In doing so, I establish historical precedence for these bias-variance tradeoffs, even though I am the first, to my knowledge, to explicitly connect the tradeoff to these debates.

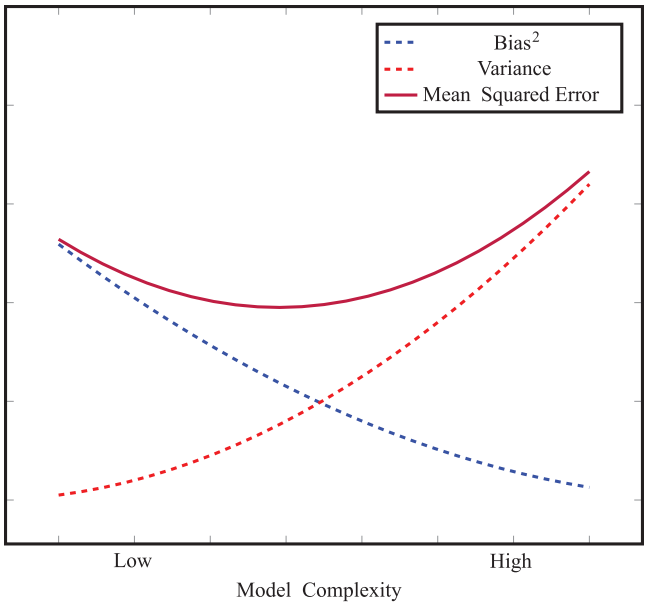


FIGURE 3. *A depiction of how bias, variance, and mean squared error change as a function of model complexity. While a bias-variance tradeoff is present, there is usually an optimal point of model complexity that minimizes mean squared error by effectively balancing bias and variance. Note that the plots shown here are completely hypothetical and do not correspond to any real data or algorithms.*

TABLE 1
Generalized Bias-Variance Decomposition Applied to Machine Learning

Machine learning	
Target T	Function f
Approximator $\hat{T}$	Estimator $\hat{f}$
Mechanism $\mathcal{M}$	Machine learning algorithm
Source of randomness	Data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $(\mathbf{x}_i, y_i) \sim \mathcal{P}_D$ for $i = 1, \ldots, n$

I note that a lot of nuance can be lost when these debates are dichotomized. Just as machine learning approaches have varying levels of bias and variance, positions on these debates are not always on one extreme in a dichotomy. However, by treating approaches as being relatively "high bias" and "high variance" (as is often done in machine learning), I hope to show how the bias-variance tradeoff can bring insights to these debates. Table 2 shows a set of dichotomies that arise in the ensuing discussion and that exhibit such tradeoffs.

Theories of Learning: Cognitivism Versus Situativism/ Constructivism

Over the past few decades, there have been several debates around the nature of learning. One prominent debate has been between cognitivists and situativists, as epitomized by the discourse between Anderson et al. (1996, 1997) and Greeno (1997). Cognitivists, or information-processing psychologists, posit that learning happens in the individual's mind and that we can develop precise theories of cognition and learning, including computational architectures and simulatable models. Situative theorists, on the other hand, claim that cognition happens in a particular sociocultural context and is thus context-dependent (J. S. Brown et al., 1989; Lave & Wenger, 1991). As such, situative theories typically do not employ precise computational models to predict learning; rather they rely extensively on qualitative techniques such as ethnography and ethnomethodology (Greeno, 1997).

TABLE 2
Dichotomies Discussed or Mentioned in This Article That Exhibit a Sort of Bias-Variance Tradeoff

Bias	Variance
(Opposite of) accuracy	(Opposite of) precision
(Opposite of) validity	(Opposite of) reliability
Underfitting	Overfitting
Theory-driven	Data-driven
Top-down	Bottom-up
Logical	Statistical
Symbolic	Connectionist
Neat	Scruffy
Cognitivism	Situativism
Cognitivism	Constructivism
Quantitative	Qualitative
Controlled experiments	Design experiments
Reductionism	Holism
Elemental	Systemic
Direct instruction	Discovery learning
Positivism	Constructivism / interpretivism

Note. Many of these phenomena are related/overlapping, but they are not synonymous.

Another related debate on the nature of learning is between cognitivists and (radical) constructivists. Radical constructivism is rooted in Piaget's genetic epistemology, and claims that every individual necessarily constructs their own reality as they learn, and there is no way of knowing if that reality corresponds to an external reality (von Glasersfeld, 1991). Constructivism and situativism may seem unrelated at first glance, but the same information-processing psychologists found themselves engaging in similar debates with both groups. Indeed, according to Anderson et al. (1998),

> The alliance between situated learning and radical constructivism is somewhat peculiar, as situated learning emphasizes that knowledge is maintained in the external, social world; constructivism argues that knowledge resides in an individual's internal state, perhaps

TABLE 3
Generalized Bias-Variance Decomposition Applied to the Debate on Educational Learning Theories

Educational learning theories	
Target T	True scientific theory of learning
Approximator $\hat{T}$	Proposed learning theory
Mechanism $\mathcal{M}$	Interpretation *and* collection of data
Source of Randomness	Data collected from human subjects (e.g., through controlled experiment, think-aloud protocol, or ethnographic study)

> unknowable to anyone else. However, both schools share the general philosophical position that knowledge cannot be decomposed or decontextualized for purposes of either research or instruction. (p. 235)

Sometimes, the debate between cognitive and situative theories is recast as being about unit of analysis—cognitivists focusing on the individual and situativists focusing on social groups and communities of individuals. In this sense, constructivism, which also focuses on individual cognition, can be recast as a flavor of cognitivism (Derry, 1996). However, this can obscure the nature of these debates. The bias-variance tradeoff can help demonstrate why situativists and constructivists (even if they differ in their unit of analysis) often find themselves agreeing, and on the opposite side of cognitivists.

Table 3 shows how the generalized bias-variance decomposition can be mapped to these debates around learning. The target is the "true" scientific theory of learning, and our goal, as learning scientists, is to find a theory of learning that can approximate the true theory. (Some learning scientists may deny that there is such a "true" theory of learning or whether our goal is really to approximate such a theory. Indeed, these epistemological differences play an important role in the educational debates we are discussing here, and we will revisit them later in the Discussion section later.) The development of a theory requires collecting data from human subjects and, through various kinds of data analysis, creating models of how people learn. Cognitivists develop precise computational models that can be simulated and fit to data—much like machine learning models. Situativists and constructivists, on the other hand, tend to rely on qualitative methods and case studies to construct qualitative models and theories about how people learn.

Situative theories developed in reaction to the cognitive tradition, with many early situative researchers originally being cognitivists. These researchers were disheartened with cognitive theories, because they could not describe the rich kinds of learning that happen in authentic environments (outside of labs and even formal schooling environments). In other words, they found cognitive theories to be biased. On the other hand, cognitive theorists found the situative perspective to be too imprecise, informal, and not generalizable, or in other words, high variance. Indeed, according to Anderson et al. (1997),

> We have sometimes declined to use situated language (what Patel, 1992, called "situa-babel") because we do not find it a precise vehicle for what we want to say. In reading the literature of situated learning, we often experience difficulty in finding consistent and objective definitions of key terms. (pp. 18–19)

Moreover, in a critique of information-processing psychology from a constructivist perspective, Cobb (1990) noted the presence of a bias-variance tradeoff[5]:

> It can be argued that there is a trade-off between accounting for the subjective experience of doing mathematics and the precision inherent in expressing models in the syntax of computer formalisms . . . [constructivism] gives greater weight to mathematical experience and [cognitivism] greater weight to precision. (p. 68)

If learning really is context-dependent, then how can the result of an ethnographic study in a very specific context generalize? Any specific study might "overfit" to a particular situation, but perhaps on average, situative theories are more accurate (less biased) because they describe learning in more complete terms. For example, situativist accounts are more likely to account for the social aspects of learning or nuances that appear only when the rich classroom context or authentic informal learning context is taken into account. On the other hand, information-processing theories are often developed by studying adults in lab settings, where the nuances that appear in authentic learning settings are minimized. Appendix B gives quotations to illustrate why constructivists and qualitative researchers view the high-variance nature of their theories essential.

To make the application of the bias-variance tradeoff here more concrete, suppose for the sake of argument that we are trying to learn a function f to predict some learning event y. For a cognitivist, the inputs of this function could be a variety of variables such as the student's knowledge, working memory, perception, cognitive load, and so on; we can call these individual variables as they represent learning in the individual's head. A situativist might also include variables such as the context in which the learning takes place, factors pertaining to other individuals involved in the learning situation, and so on. According to Greeno (1997, 1998), cognitivists make a "factoring assumption," whereby they assume that cognition can be nearly decomposed into individual factors (such as the inputs mentioned above) that are largely independent of one another. For Greeno, an important oversight of this factoring assumption is that the context does not affect the role that the other variables play. In other words, the factoring assumption assumes the function can be decomposed as follows:

$$y = f(\text{individual variables, context variables}) \approx g(\text{individual variables}) + h(\text{context variables}),$$

and in many cases, $h(\text{context variables}) = 0$ (i.e., context may be ignored). However, for Greeno and other situativists, such a decomposition is not possible in many cases as variables pertaining to the individual will interact with variables pertaining to the environment, including other people in the environment. The difference here can be now be viewed in terms of linear regression. A regression with no interaction terms is lower variance (but often higher bias, assuming interactions do exist) than a regression that includes interaction terms. The more interactions, the higher the variance. Of course, interactions can be modeled in more complex ways. For example, in an extreme case, the dependency of f on the individual variables could change drastically as the context variables change (e.g., a student's ability to utilize some piece of mathematical knowledge may be highly dependent on whether the student is at school, at home, in the kitchen, in the marketplace, etc.). The factoring assumption also plays a role in constructivist critiques of cognitivism (Shepard, 1991); see Appendix C for a concrete example of how it could account for the bias-variance tradeoff in this debate.

The bias-variance tradeoff described here can also explain debates in methodology that appear in education research, and the social sciences more generally (i.e., quantitative vs. qualitative, reductionistic vs. holistic, and controlled experiments vs. design experiments); the connection to these debates is briefly described in Appendix D.

Analogues to Bias-Variance Tradeoffs in Artificial Intelligence. The bias-variance tradeoff in learning theories can be better understood by noticing that it actually has historical roots in similar tradeoffs that have appeared in the history of AI. The cognitive perspective is rooted in the tradition of symbolic and rule-based approaches that were popular in the early days of AI research. Rule-based AI is often seen as being biased because it is theory-driven, in contrast to more data-driven approaches to AI, such as machine learning, especially connectionist neural networks.[6] On the other hand, neural networks and other black box machine learning algorithms have higher variance because they are more susceptible to overfitting to particular data. While connectionism seemingly has little to do with educational theories of learning, parallels have been drawn between connectionism, situativism, and constructivism (Quartz, 1999; Vera & Simon, 1993; Winograd, 2006); indeed, neural networks could help do away with the factoring assumption mentioned earlier. For example, diSessa (1993) proposed an influential constructivist theory of how students develop intuitive conceptions in fields such as physics through the activation of networks of p-prims ("phenomenological primitives"). While this model was developed by diSessa and utilized by later researchers using deep qualitative interviews of students, diSessa (1993) also initially sketched a connectionist architecture for how p-prims might form networks that could result in conceptual change (though such a connectionist architecture was never implemented to my knowledge). In this sense, the bias-variance tradeoff between educational theories of learning is not so far removed from the associated tradeoff between theory-driven and data-driven approaches in AI and machine learning.[7]

However, in the history of AI, another distinction of approaches was made by Schank (1983) that is more useful for our purposes: the distinction between "neats" and "scruffies." According to Kolodner (2002),

> "Neats" . . . were taking a careful, experimental, and (to us "scruffies") slow route toward getting small results that would hopefully add up to a coherent big picture. "Scruffies," on the other hand, were taking a more intuitive and holistic and (to the "neats") far messier approach, using modeling and observational methods to get at the big picture, hoping that it will provide guidance on which smaller details to focus on. While neats focused on the way isolated components of cognition worked, scruffies hoped to uncover the interactions between those components. (p. 139)

It should be clear where the bias-variance tradeoff lies in these approaches. Kolodner (2002) pointed out that Herbert Simon, Allen Newell, and John Anderson were prototypical neats, while Seymour Papert, Marvin Minsky, and Roger Schank were prototypical scruffies. Many of these names should be recognizable. Interestingly, all of these researchers not only were AI and cognitive science researchers but also played an important role in defining theories that would affect education research. Simon, Newell, and Anderson were three of the pioneers of information-processing psychology. Papert was a student of Piaget (one of the founders of constructivism), and he and Minsky developed an approach to AI and education research that drew heavily on constructivist ideas. Papert later founded constructionism, a learning theory that built on Piaget's constructivism (Papert, 1988). Schank, who coined the term *learning sciences* and effectively founded the field (Schank, 2016), took a similar approach to Papert and Minsky in his AI research (Brockman, 1996) and education research. Similar tradeoffs can be seen in other areas of cognitive science. For example, in linguistics, Noam Chomsky can be regarded as a neat whose Universal Grammar is acknowledged as being a high-bias (low-variance) theory (Lappin & Shieber, 2007; Norvig, 2017). Indeed, Chomsky engaged in debates with both Piaget and Schank, and more recently, researchers have used higher variance machine learning models and computational learning theory to challenge his theory (Lappin & Shieber, 2007). Thus, the existence of a bias-variance tradeoff in learning theories is not a coincidence but rather an offshoot of similar tradeoffs in AI, which gave rise to the field of machine learning, where the tradeoff was later formalized.

TABLE 4
Generalized Bias-Variance Decomposition Applied to the Pedagogical Debate on Direct Instruction Versus Discovery Learning

Pedagogy	
Target T	Optimal educational experience for each student
Approximator $\hat{T}$	Educational experience that each student actually receives
Mechanism $\mathcal{M}$	Instructional intervention
Source of randomness	Stochasticity in what students do over the course of the instructional intervention

Pedagogy: Direct Instruction Versus Discovery Learning

The debate around direct instruction versus discovery learning has surfaced in various ways throughout the history of education (Bruner, 1961; Ray, 1961; Winch, 1913), and is still a topic of controversy in education research (Kirschner et al., 2006; Munter et al., 2015; Tobias & Duffy, 2009). *Direct instruction* refers to directly and explicitly teaching students what you want them to learn. *Discovery learning* suggests that students should be left to discover and construct knowledge for themselves with limited intervention from teachers. In many ways, direct instruction is rooted in cognitive theories and discovery learning is rooted in constructivist theories of learning (Kirschner et al., 2006), and as such, this debate parallels the debate around learning theories; the bias-variance tradeoff can help us see this relationship from another angle.

Table 4 shows a mapping of the generalized bias-variance decomposition to this debate on pedagogy. Here the goal is to come up with the best educational experience for each student (i.e., a function that maps educational experiences to students). The instructor must choose an instructional intervention, and that intervention will in turn determine the precise educational experience that each student receives. Notice that the process by which the intervention results in an educational experience for each student is assumed to be stochastic because it depends on how the intervention actually pans out, how receptive the student is to the intervention, various unrelated distractions that might come up in the classroom, and so on. [8] For example, the teacher may start with some lesson plan, but if a student appears to be struggling, the teacher may change the lesson or instructional technique as needed. On the other hand, the teacher may decide to give students time to research an open-ended project, therefore giving the students much control over their educational experiences, resulting in varied experiences depending on each student's level of motivation, prior knowledge, metacognitive skills, and so on. Notice that the source of randomness here is very different from randomness in the data used to create a theory or fit a machine learning model. That is, here the instructor is not using randomly generated data to construct an instructional strategy. Rather, the instructor can predetermine their instructional strategy ahead of time, but the result of that strategy is not completely in the instructor's hands.

The bias in an instructional strategy is the average degree to which the educational experience that a student receives is suboptimal. The variance in an instructional strategy is the variance in educational experiences that a student might receive as a result of that strategy. Direct instruction falls on the bias side of the spectrum because students are directly taught what the instructor or other experts *believe* the students need to know, which may not actually be what each student would benefit from most. For example, Spiro et al. (1988) explicitly pointed out several "reductive biases" that could result from direct instruction in ill-structured domains, whereby students oversimplify the complexities and nuances of the domain.

On the other hand, discovery learning lies on the variance side of the spectrum, as (in its most extreme form) it suggests leaving students to discover the best educational experience for themselves. As such, there is a lot of variation in the possible educational experiences that students end up receiving. Proponents of direct instruction would cite this as a limitation of discovery learning; students might end up reaching a dead end in their discovery process or construct misconceptions. Proponents of discovery learning find direct instruction to be overly focused on direct knowledge acquisition and performance-based metrics in well-defined domains but insufficient to account for richer notions of learning such as learning in ill-defined domains (Spiro et al., 1988), preparing for future learning (Schwartz et al., 2009), or becoming a participant in a community of practice (Lave & Wenger, 1991). These other forms of learning naturally require varied educational experiences to support the variety of situations in which students might have to use (or reconstruct) their knowledge in the future.

Papert (1987b) explicitly discussed how high variance is necessary in order to allow for different students to receive the right educational experience for them (in the context of children learning mathematics through the Logo programming language):

> Whenever children are exposed to this sort of thing, a certain number of children seem to get caught by discovering zero. Others get excited about other things.
>
> The fact that not every child discovers zero this way reflects an essential property of the learning process. No two people follow the same path of learnings, discoveries, and revelations. You learn in the deepest way when something happens that makes you fall in love with a particular piece of knowledge. (p. 82)

Navigating the Bias-Variance Tradeoff

While the bias-variance tradeoff seems to imply that by reducing bias one has to increase variance or vice versa, such a tradeoff is not necessarily an equal exchange of bias and variance, meaning there could be ways to find an "optimal" amount of bias and variance, as depicted by the minimum of the mean squared error plot in Figure 3. In the machine learning literature, there are many proposed techniques for effectively navigating the bias-variance tradeoff in order to minimize mean squared error. Indeed, throughout the history of AI, researchers have advocated for combining symbolic or logical AI approaches with (high variance) statistical AI (Bach et al., 2017; Domingos et al., 2006; Hu et al., 2016; Minsky, 1991). For example, in an article called "Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy," Minsky (1991) argued that

> neither purely connectionist nor purely symbolic systems seem to be able to support the sorts of intellectual performances we take for granted even in young children. . . . I'll argue that the solution lies somewhere between these two extremes, and our problem will be to find out how to build a suitable bridge. (p. 37)

We now turn to some concrete techniques for effectively navigating the bias-variance tradeoff to minimize mean squared error in the educational debates we have discussed.

Increasing the Amount of "Data"

As mentioned before, as the amount of data increases, the variance of a machine learning function class decreases. Thus, as the amount of data increases, higher variance techniques become more effective. As the amount of data goes to infinity, then the best function class is one that has zero bias, no matter what the variance is (since it will diminish). This can also be seen in terms of reliability and validity in social sciences research. It is okay to have a high-validity, low-reliability survey if the number of survey respondents is very large. One consequence of this is that one should pick a method with the right amount of variance for the amount of data one has. Another implication is that if one wants to use a high-variance method (e.g., because it has less bias than other methods), one should collect more data to minimize overfitting. Of course, collecting more data is not always easy; for example, conducting ethnographic studies of many learning situations can be quite costly, which can limit one's ability to create a generalizable theory of situated learning (if that is one's goal).

But how does "more data mean less variance" give us insights into debates around pedagogy, where data are not present? We can think of data as a limited resource that regulates the degree of stochasticity in machine learning algorithms. An analogous resource in the case of debates around pedagogy is instructional time. Much of the argument against discovery learning is around efficiency. While it would be great if students could rederive all scientific laws, who has the time to do that? It took centuries the first time around! Direct instruction is efficient, and if time is limited, it can lead to more quickly disseminating knowledge to students. But if one has more instructional time available, then perhaps some of that time could be spent on discovery activities where students can develop more robust understandings. Indeed, in Mehta and Fine's (2019) detailed study of potentially promising American high schools, the authors found that most schools and teachers were not able to provide deeper learning opportunities to their students due to the need to cover a lot of instructional material, which is done more efficiently via lecture; however, one progressive project-based learning school was able to effectively engage students in sustained deeper learning by largely avoiding the demands of traditional standards.

Raw instructional time is not the only resource of interest; sometimes a more relevant variable is the number of instructional activities of interest. Spiro et al. (1991) argued that for ill-structured domains where there is "case-to-case irregularity" students need to see a variety of cases in order to have robust knowledge of that domain. Not only does direct instruction not suffice due to its proneness to "oversimplification" and other "reductive biases," but the more examples students see, the more likely they are to generalize better. For example, medical students would need to encounter several cases of different patients with similar conditions due to the nuances that appear across cases; simply reading books about medical conditions is not enough. But if instructional time is limited, teaching generic principles via direct instruction might be more effective than having students work with one or two cases, as they might "overfit" their understanding to those particular cases. Therefore, discovery learning is useful in this case only if a student can see sufficiently many cases in order to obtain a generalizable understanding of the ill-structured domain.

Regularization

In machine learning, regularization is one of the most common techniques to mitigate variance (at the expense of adding some bias). Rather than simply trying to only minimize mean squared error, one adds a regularization parameter that constrains the complexity of the function class. In particular, the objective of a machine learning algorithm might be as follows:

$$\min_{\hat{f}\in\mathcal{F}} \sum_{(\mathbf{x},y)\in D} (\hat{f}(\mathbf{x})-y)^2 + \lambda * \text{Complexity}(\hat{f}).$$

The term on the left is the empirical mean squared error on the data set. If one just minimizes that quantity, one risks overfitting (especially with a high-variance function class) to the particular data set. The left-hand term is some measure of the complexity of $\hat{f}$ and λ is a regularization parameter.

The higher λ is, the more a function is penalized for being too complex. This effectively reduces the size or complexity of a function class by penalizing functions that might overfit to data. Regularization could add a small amount of bias but with the hopes of greatly decreasing variance.

Analogously, in teaching, discovery learning can be made more effective by adding in a small amount of guidance, which may prevent the student from getting lost in their discovery process at the expense of biasing the student towards the teacher's "right answer." This guidance could bias students away from their optimal educational experience but may typically lead to a more productive experience than letting students figure everything out on their own. Indeed, researchers on both sides of the direct instruction versus discovery learning debate have realized that the optimal form of instruction is somewhere in the middle, often referred to as *guided* discovery learning (A. L. Brown & Campione, 1994; Gagné & Brown, 1961; Kersh & Wittrock, 1962; Tobias & Duffy, 2009). Like a regularization parameter, a "guidance parameter" can regulate how much a student should struggle before receiving guidance. A small amount of struggle and failure could be productive, but letting students drown in the deep end of discovery is not. The question then becomes how much guidance is optimal. In machine learning, to find the optimal regularization parameter, one can use a technique known as cross-validation; there is no clear analogue to cross-validation in the instructional setting, but a similar procedure would simply be to try different amounts of guidance to learn over time how much guidance seems to be optimal. Of course, the optimal amount of guidance might vary from student to student and from one instructional situation to another. Perhaps, in some cases it is best for a good teacher to rely on their intuition (which is not altogether uncommon in setting regularization parameters either).

Model Ensembles

Model ensemble learning techniques, such as bagging, boosting, and stacking, combine multiple models to reduce the bias and/or variance of the individual models when making predictions. For example, stacking or stacked generalization can take multiple models as input and then use a meta-algorithm (e.g., linear regression) to assign a weight to each subalgorithm to make a final prediction that could correct for biases in the input models (Breiman, 1996; Wolpert, 1992).

The main takeaway for our purposes is that by combining multiple models/theories, we can perhaps mitigate the biases or variance in the individual models. For example, by looking to various studies of situated learning or learning in constructivist classrooms, we can look for overarching trends and patterns that appear across the studies. Such trends are likely to not overfit to particular situations but rather could give us insights that might generalize well across situations. At the same time, these insights might avoid biases in cognitive models that disregard richness and social dynamics of authentic learning environments.

To see how model ensembles can help mitigate bias, we can take inspiration from Papert and Minsky's approach to AI. In an early report on their state of AI research, Minsky and Papert (1971) discussed their use of *microworlds* in contrast to rigid symbolic approaches to AI (like that of Simon, Newell, and Anderson):

> We are dependent on having simple but highly developed models of many phenomena. Each model—or "micro-world" as we shall call it—is very schematic . . . we talk about a fairyland in which things are so simplified that almost every statement about them would be literally false if asserted about the real world. Nevertheless, we feel they are so important that we plan to assign a large portion of our effort to developing a collection of these micro-worlds and finding how to embed their suggestive and predictive powers in larger systems without being misled by their incompatibility with literal truth.

In other words, they were comfortable with using a collection of models that were admittedly very biased, because perhaps by integrating these biased models together they could create less biased systems (possibly at the expense of being higher variance). Papert (1980, 1987a) extended this AI notion of microworlds to "slices of reality" that children can interact with (e.g., on a computer) when learning. Each of these microworlds allow students to discover and explore but with bounds that constrain the space of discovery, bounds that bias the world potentially into being very incomplete and inaccurate conceptions of reality. But to Papert this did not matter; it is by discovering reality through multiple (biased) microworlds, that a learner could develop robust understanding of the macroworld (see Appendix B for a brief exposition of microworlds in Papert's educational theory).

This is similar to Spiro et al.'s (1988) suggestion of needing multiple cases to learn about an ill-structured domain. In fact, (Spiro et al., 1991) advocated for the need to be able "to construct from those different conceptual and case representations a knowledge *ensemble* tailored to the needs of the understanding or problem-solving situation at hand" (p. 24). While each case is biased and not wholly representative of the domain at large, by examining many different cases, a learner can develop more robust knowledge about the domain. While these ideas suggest high-level parallels between ensemble learning and approaches to discovery learning, perhaps more formal connections can be made to existing ensemble learning techniques.

Discussion

Here we discuss two other important differences that cut across the various bias-variance tradeoffs that we have

previously discussed. Namely, we look at (1) the distinction between predictive (or descriptive) power and prescriptive power and (2) how these bias-variance tradeoffs relate to debates around epistemology, which can help explain why these debates linger to this day.

Predictive Power Versus Prescriptive Power

Another difference that can arise between techniques that tend to have more bias versus techniques that tend to have more variance is that the former are often more conservative (often relying on existing theories), which in turn is better at *predicting things as they are*. On the other hand, higher variance approaches tend to explore a broader space of potential solutions, which while prone to overfitting could lead to *prescribing new solutions*.

This point was made clear by Papert (1980) when contrasting Minsky's and his approach to AI and education research with that of other AI researchers like Newell and Simon. He saw a key difference in their approach being "seeing ideas from computer science not only as *instruments of explanation* of how learning and thinking in fact do work, but also as *instruments of change* that might alter, and possibly improve, the way people learn and think" (pp. 208–209).

To contextualize this claim, let us first examine Simon and Newell's (1971) approach to AI. Their interest was in developing AI programs for problem-solving that that were aligned with how humans solve problems, which in turn would also lead to their cognitivist theory of information-processing psychology. In fact, their approach was to develop programs intentionally limited by human capabilities (see Appendix B). This is in stark contrast to modern machine learning approaches that are primarily data-driven and allow for computational techniques that humans very likely do not use (at least not consciously) when solving problems. This allows for developing programs that can detect patterns in high-dimensional data that people can simply not do. More recently, deep learning has demonstrated the ability to create AI that can exhibit superhuman performance. Recent deep learning agents such as DeepMind's AlphaGo (an agent that can beat experts at the ancient game of Go; Silver et al., 2016) and AlphaStar (an agent that has beaten experts at the popular computer game StarCraft II; Vinyals et al., 2019) not only beat humans but also do so by exhibiting new strategies to playing these games that humans can study and possibly learn from (Chan, 2017).

Similarly, Papert and Minsky were interested in finding ways to change how people learn altogether. That is, while Simon, Newell, and other cognitivist researchers were interested in more efficiently teaching students by understanding how the mind works; Papert and Minsky were interested in finding ways to fundamentally alter and improve how people learn by understanding how the mind *could* work.

For example, while a firm student and proponent of Piaget, Papert saw his mentor's theory of developmental stages as being a *predictive* theory, but not a *prescriptive theory*. According to Papert (1987b), "One might say that the formal stage arrived so late precisely because there were no computers" (p. 93). Taking a more discovery-oriented approach, Papert (1980, 1987b) believed that by using microworlds, children could reach the formal operational stage before researchers like Piaget had ever observed previously. Thus high-variance approaches do not find themselves confined by existing theories, and hence can find solutions outside of the scope of such theories, which may have a biased way of looking at things as they are. However, it is important to note that Papert's starting point was Piagetian theory; it is merely by tweaking the theory that he was able to escape it. This points again to the need for techniques that can combine the best worlds, for example, by taking the predictive power of existing theories as a starting point to establish powerful prescriptive theories and instructional techniques.

Epistemology Debates

If the solution is to combine aspects from both sides of a debate (e.g., cognitivist and situativist theories, or discovery and direct instruction), and researchers realize that moderate positions are better than the extremes, then why do these debates persist? It may seem as though pragmatic considerations should determine the optimal mix of different positions (and the optimal amount of bias and variance), but I claim that these debates linger on, at least in part, due to different philosophical worldviews and epistemologies.

Advocates of cognitive theories, quantitative methodologies, and direct instruction *tend* to have realist and positivist (or postpositivist) beliefs about the nature of knowledge, learning, and the world. That is, they believe there is a "true" theory out there and empirical observations can help us approximate, or even discover, that truth. On the other hand, advocates for situative theories, qualitative methods, and discovery learning tend to hold constructivist[9] and interpretivist epistemologies, positing that each person's understanding of the world is a mental construction and does not necessarily correspond to an external reality (which does not mean they necessarily deny the existence of such a reality; von Glasersfeld, 1991).

A positivist researcher is likely to believe their theory is correct or at least an approximation of reality, and thus may not readily admit that it is biased. For similar reasons, realists are also more likely to be fond of teaching an "essential" standardized curriculum to all students, not viewing their curriculum as biased by their understandings of reality. Constructivist researchers, on the other hand, are comfortable admitting that they and their colleagues will each have different theories of learning based on their experiences,

none of which are *correct* but all of which might give insights onto reality; constructivists accept that their theories are naturally high variance. Moreover, constructivist teachers believe that their students will construct different understandings, and thus higher variance instructional strategies are necessary to suit individual differences; thus, constructivist educators will yield some of their authority in the classroom to students as they construct the subject matter for themselves (Munter et al., 2015; von Glasersfeld, 1991).

Speaking in terms of the target diagrams, an epistemologically constructivist researcher or educator may think that (1) there is no target (objective reality) to begin with or (2) it is a movable target that will vary from researcher to researcher, from context to context, or from student to student. The goal may not be to pinpoint one true universal theory of learning or to adopt one universally applicable teaching practice, but to find the approach that best fits the problem here and now. When viewed in the context of generalization however; this results in high variance.

Many social scientists have argued that these epistemological differences can amount to an incommensurability in different research paradigms (Lincoln et al., 2017), adopting the concept from Thomas Kuhn's (1962) depiction paradigms in the natural sciences. However, some have argued that incommensurable (even as initially intended by Kuhn) does not imply that different paradigms cannot be meaningfully compared in order to arrive at a richer understanding of the various perspectives (Bernstein, 1983; Cobb, 2007; Donmoyer, 2006). Indeed, a recent trend appears to be that many education researchers are explicitly choosing to adopt *pragmatism* (in the tradition of John Dewey, Charles Sanders Peirce, and William James) as their epistemological stance to move away from age-old epistemological debates that might hinder the progress of doing good education research (Cobb, 2007; Doroudi et al., 2020; Taber, 2010). Perhaps because constructivists are accepting of the limitations to their approach (i.e., that it is high variance), they tend to be more pragmatic and are willing to draw from competing approaches to navigate the bias-variance tradeoff (Cobb, 2007; Danish & Gresalfi, 2018; Derry, 1996; Greeno, 1998; Sfard, 1998; Shaffer, 2017). On the other hand, if a realist/positivist researcher or teacher denies that their approach is biased, then they may not see the need to draw from researchers who adopt alternative epistemologies (Anderson et al., 1997).

It is important to note that these different epistemologies are not necessary aspects of the bias-variance tradeoffs. For example, in machine learning, many researchers will tend to adopt a positivist or empiricist approach when choosing to find the model that best fits the data, regardless of where the approach lies on the bias-variance tradeoff. This could also explain why the use of neural networks (in a traditional machine learning fashion) does not have a closer historical correspondence to the research approach of constructivists or situativists; they do not align epistemologically.

Conclusion

I have shown how the bias-variance tradeoff in machine learning can be formally generalized to provide insights into age-old educational debates. While this framework does not resolve these debates, it can justify why the different positions are all trying to do something meaningful (in terms of "minimizing mean squared error"). It can also help explain relationships between concepts that may seem unrelated at first glance. For example, while situativist and constructivist theories sometimes have different units of analysis, they are aligned by their position on the bias-variance tradeoff. Similarly, due to their high-variance nature, neural networks can be seen as being related to situativist and constructivist theories, although that connection is limited due to epistemological differences.

Moreover, by looking toward concrete data science and machine learning techniques to navigate the bias-variance tradeoff, we can discover analogous ways of tackling similar tradeoffs in education. Many of these ways of navigating the tradeoffs have been proposed throughout the history of AI and the learning sciences, but sometimes only in passing in isolated articles, rather than being presented systematically. By situating these debates onto the bias-variance tradeoff, we can systematically identify solutions that have been proposed in the past as well as offer new solutions by looking to the machine learning literature. This article simply begins to illustrate how certain concepts such as the bias-variance tradeoff, overfitting, regularization, and model ensembles relate to approaches in education research, with the hopes of inspiring researchers to find ways to concretely and productively apply these concepts to educational problems. Perhaps by working together, machine learning researchers and education researchers can find new ways to utilize the science of data to productively advance the science of teaching and learning in ways that move past seemingly never-ending paradigmatic differences.

Appendix A

Proof of Generalized Bias-Variance Decomposition

Theorem 2 (Generalized Bias-Variance Decomposition). *Suppose our goal is to approximate some target*: $T : \mathbb{R}^m \to \mathbb{R}^n$. *Let* $\mathcal{M}$ *be a stochastic mechanism (i.e., a partially random process) that randomly chooses a function* $\hat{T} : \mathbb{R}^m \to \mathbb{R}^n$ *from a function class* $\mathcal{T}$ *(i.e.,* $\hat{T} \sim \mathcal{M}$*). Then we have the following for all* $\mathbf{x} \in \mathbb{R}^m$ *:*

$$\underbrace{\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - T(\mathbf{x}))^2]}_{\text{Mean Squared Error}} = \underbrace{(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))^2}_{\text{Bias Squared}} + \underbrace{\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2]}_{\text{Variance}}.$$

Proof. By adding and subtracting $\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})]$ we have the following:

$$\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - T(\mathbf{x}))^2] = \mathbb{E}_{\mathcal{M}}\begin{bmatrix}(\hat{T}(\mathbf{x}) - T(\mathbf{x}) + \\ \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2\end{bmatrix}.$$

Rearranging terms, we get the following:

$$\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - T(\mathbf{x}))^2] = \mathbb{E}_{\mathcal{M}}\begin{bmatrix}(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}) + \\ \hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2\end{bmatrix}.$$

By distributing the square, the right-hand side can be written as follows:

$$\mathbb{E}_{\mathcal{M}}\begin{bmatrix}(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))^2 + (\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2) \\ + 2(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})]\end{bmatrix}.$$

By the linearity of expectation, the right-hand side can be written as follows:

$$\mathbb{E}_{\mathcal{M}}\left[(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))^2\right] + \mathbb{E}_{\mathcal{M}}\left[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2\right] + \mathbb{E}_{\mathcal{M}}\left[2(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])\right].$$

Notice that $\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x})$ is not a random variable. Thus, the right-hand side can be re-written as follows:

$$(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))^2 + \mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2] + 2(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])].$$

Finally, notice that $\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])] = 0$. This gives us the desired decomposition:

$$\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - T(\mathbf{x}))^2] = \underbrace{(\mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})] - T(\mathbf{x}))^2}_{\text{BiasSquared}} + \underbrace{\mathbb{E}_{\mathcal{M}}[(\hat{T}(\mathbf{x}) - \mathbb{E}_{\mathcal{M}}[\hat{T}(\mathbf{x})])^2]}_{\text{Variance}}.$$

Appendix B

Representative Quotations

The Necessity of Variance in Qualitative Research. Papert (1987b), a constructivist, explicitly mentioned advantages of deep qualitative studies of small samples in order to have a develop a less biased understanding of how people learn:

> One of our students at MIT, Robert Lawler, wrote a Ph.D. thesis years ago based on his observation of a six-year-old child. Over a period of six months, he observed this child almost continuously, never missing as much as a half hour.[10] . . . When people study the learning process, they usually study a hundred children for several hours each, and Lawler showed very conclusively . . . that you lose a lot of very important information that way. By being around all the time, he saw things with this child that he certainly would never have caught from occasional samplings in the laboratory. (p. 90)

In the broader context of using "thick description" in qualitative research, Geertz (1973) made a similar claim about the importance of accepting variance in coming up with a scientific theory that generalizes:

> It is, in fact, by its power to draw general propositions out of particular phenomena that a scientific theory—indeed, science itself—is to be judged. If we want to discover what man amounts to, we can only find it in what men are: and what men are, above all other things, is various. It is in understanding that variousness—its range, its nature, its basis, and its implications—that we shall come to construct a concept of human nature that, more than a statistical shadow and less than a primitivist dream, has both substance and truth. (p. 58)

Model Ensembles and Microworlds. Analogous to his use of microworlds in AI, Papert (1987a) saw microworlds as things children naturally construct in their heads, to make sense of the world around them:

> Probably in all important learning, an essential and central mechanism is to confine yourself to a little piece of reality that is simple enough to understand. It's by looking at little slices of reality at a time that you learn to understand the greater complexities of the whole world, the macroworld. (p. 81)

Although children construct microworlds on their own, Papert believed the process could be greatly facilitated by giving them physical or digital microworlds to work with. He and his colleagues created such microworlds in the Logo programming language (and its variants). According to Papert (1980), students can learn better by interacting with different kinds of microworlds to overcome the particular biases of any one microworld:

> So, we design microworlds that exemplify not only the "correct" Newtonian ideas, but many others as well: the historically and psychologically important Aristotelian ones, the more complex Einsteinian ones, and even a "generalized law-of-motion world" that acts as a framework for an infinite variety of laws of motion that individuals can invent for themselves. Thus learners can progress from Aristotle to Newton and even to Einstein via as many intermediate worlds as they wish. (p. 125)

Having a firm understanding of gravity might be difficult if we have only experienced one kind of gravity ($-9.8\ \mathrm{m/s^2}$) our entire lives. By experiencing different kinds of gravity—or being able to manipulate gravity and see the effects for ourselves—we can perhaps begin to form a more accurate concept of it. Papert (1980) explained why it might be useful to construct "incorrect" microworlds by recalling examples of this that are developmentally important:

> This is an effective way to learn, paralleling the way in which each of us once did some of our most effective learning. Piaget has demonstrated that children learn fundamental mathematical ideas by first building their own, very much different (for example, preconservationist) mathematics. And children learn language by first learning their own ("baby-talk") dialects. So, when we think of microworlds as incubators for powerful ideas, we are trying to draw

> upon this effective strategy: We allow learners to learn the "official" physics by allowing them the freedom to invent many that will work in as many invented worlds. (pp. 126–127)

Predictive Power Versus Presecriptive Power. An illustrative example that contrasts Simon and Newell's approach from both Papert and Minsky's, as well as modern-day AI (especially deep learning), is given in one of their early publications on their strategy toward developing AI programs. According to Simon and Newell (1971), one of the steps of this strategy was as follows:

> Discover and define a program, written in the language of information processes, that is capable of solving some class of problems that humans find difficult. Use whatever evidence is available to incorporate in the program processes *that resemble those used by humans.* (*Do not admit processes, like very rapid arithmetic, that humans are known to be incapable of*). (p. 146, emphasis added)

While Papert and Minsky were also interested in learning how humans (especially children) think and learn, they did not shy away from the possibility that the status quo could be fundamentally changed. Their focus on children as compared to Simon and Newell's focus on adults is also illustrative. Papert and Minsky were interested in development, while Simon and Newell were primarily interested in performance (seeing learning and development as more of an afterthought). According to Newell and Simon (1972),

> If performance is not well understood, it is somewhat premature to study learning. Nevertheless, we pay a price for the omission of learning, for we might otherwise draw inferences about the performance system from the fact that the system must be capable of modification through learning. It is our judgment that in the present state of the art, the study of performance must be give precedence, even if the strategy is not costless. Both learning and development must then be incorporated in integral ways in the more complete and successful theory of human information processing that will emerge at a later stage in the development of our science (p. 8).

Their later strategy for accounting for learning via production systems (Simon & Newell, 1971)—akin to expert systems in AI—was indeed a simple (high-bias) way to account for what Papert, Minsky, Schank, and others would see as a much more complex process.

Appendix C

The Factoring Assumption and Knowledge Decomposition

According to Anderson et al. (1998) and Anderson et al. (1999), a common critique of constructivism and situativism against cognitivism is that the latter assumes knowledge can be (nearly) decomposed into independent components. This hypothesis of knowledge decomposition is still prevalent to this day and underpins cognitive psychology based approaches to intelligent tutoring systems (Aleven & Koedinger, 2013; Koedinger et al., 2013). The knowledge decomposition hypothesis can be viewed as a specific case of the factoring assumption (Greeno, 1997). Suppose KC_i represents the student's knowledge of knowledge component i (e.g., as a number or a vector). Then the factoring assumption here implies that

$$y = f(\mathrm{KC}_1, \mathrm{KC}_2, \ldots, \mathrm{KC}_n) \approx f_1(\mathrm{KC}_1) + f_2(\mathrm{KC}_2) + \ldots + f_n(\mathrm{KC}_n).$$

However, a constructivist would say that knowledge is highly dependent on prior knowledge and how that knowledge is organized in an individual's head. According to Shepard (1991),

> Because we know that learning requires reorganizing and restructuring as one learns, a more organic conception is needed. In contrast to linear hierarchies, researchers now more often depict knowledge acquisition by using semantic networks that show connections in many directions. (p. 7)

More sophisticated machine learning models such as neural networks could (in theory) capture complex interdependencies between bits of knowledge at the expense of increasing variance; I explore this connection further in the section "Analogues to Bias-Variance Tradeoffs in Artificial Intelligence."

Appendix D

Methodology Debates

The preceding discussion hints at how the bias-variance tradeoff can be extended to debates around methodology, such as quantitative versus qualitative[11] (Johnson & Onwuegbuzie, 2004) and reductionistic versus holistic (or elemental vs. systemic; Nathan & Sawyer, 2014; Salomon, 1991) methodologies. In a sense, methodology can be thought of as analogous to algorithms in machine learning. For example, linear estimators form a high-bias, low-variance function class, and linear regression is the algorithm used to discover best-fitting linear estimators. Similarly, neural networks form a high-variance, low-bias function class, and back propagation is an algorithm used to fit neural networks.

However, notice that one difference with machine learning is that in education theory development, researchers who use different methods will often be examining different sources of data. For example, data from a randomized control trial, data from problem-solving sessions in a lab study, log data from an intelligent tutoring system, and ethnographic data from informal learning environments are all different from one another. In classical machine learning, typically an algorithm is chosen to fit a function to *already collected* data; the choice becomes what method to use to analyze those data, and some methods have higher bias or variance than others. In constructing learning theories, researchers must first choose what kind of data to collect (or study, if using previously collected observational data), which in turn often helps determine the most meaningful

methods to analyze the data. Thus, as noted in Table 3, the mechanism $\mathcal{M}$ that eventually determines the proposed theory is not only the choice of how to interpret the data but also the choice of what data to collect.

Another particular debate around methodology that exhibits a bias-variance tradeoff is around controlled experimentation versus design experiments (Cobb et al., 2003) or design-based research (Barab & Squire, 2004). The former tries to find out "what works" under neat controlled conditions, while the latter tries to design systems that can support learning by taking the environmental context into account, namely, understanding "'how, when, and why' it works" (Cobb et al., 2003). Since the environment is messy and high variance, design-based research approaches are necessarily iterative (Cobb et al., 2003). Winograd (2006) noted that this iterative nature is shared between how neural networks work in AI and design research in human-computer interaction, and contrasted these approaches with symbolic cognitivist approaches in both AI and human-computer interaction. Similarly, in education, cognitivists tend to advocate for experimental studies of what works (Anderson et al., 1998; Koedinger et al., 2013) while constructivist and situativist researchers tend to advocate for design-based research (Barab & Squire, 2004; Cobb et al., 2003; Papert, 1987a). However, it is important to note that many researchers pragmatically advocate for a mix of both approaches, which could involve conducting experimental studies of designed artifacts that do not try to control for all aspects of rich, classroom environments (Koedinger et al., 2013; Pea, 1987).

Acknowledgments

Winograd (2006) identified a division of AI and HCI researchers into two camps that cut across both fields, which effectively align with the bias-variance tradeoffs we see in education research as well (see Appendix D). While he did not refer to the bias-variance tradeoff explicitly, reading this work led me to the recognition of the role that the tradeoff plays in educational debates. I would also like to thank Alexander Ihler for the suggestion of thinking of instructional time as a resource that is analogous to "data" in the context of debates around pedagogy.

ORCID iD

Shayan Doroudi https://orcid.org/0000-0002-0602-1406

Notes

1. Technically, the smallest function class is to just use the class of a single function—the function that you are trying to predict (e.g., $y = 3.5x^2 - x$) but of course—that requires already knowing the target function! A more realistic small function class in this case would be the class of all quadratic functions.

2. Note that here we are defining bias, variance, and mean squared error for a particular point x. One is often interested in the average bias, variance, and mean squared error of a function class, which can be obtained by averaging over all x. This averaging does not effect the decomposition, so we do not include it for simplicity.

3. Readers with some background in machine learning might be interested to know that recent work has shown that the "classical U-shaped" curve shown in Figure 3 appears to not always hold, such as when neural networks improve in accuracy even as the number of parameters increases. This is because, counterintuitively, seemingly more complex functions could impose some structure (or "inductive bias") on the prediction task (Neyshabur et al., 2014), which could result in a decrease in the variance (Dwivedi et al., 2020). This does not violate the bias-variance tradeoff per se but indicates that there is nuance to what constitutes a good measure of "model complexity."

4. Note that this theorem could be further generalized to apply to not just functions—for example, in the case of an archer shooting at the target, the archer always aims for the bullseye, so a more appropriate is $T = (0, 0)$. In this case $T \in \mathrm{R}^2$ is just the Cartesian coordinate of where the arrow lands. However, since all of the examples we are considering can be mapped onto functions, we will restrict ourselves to functions.

5. Cobb (1990) actually used the terms *weak program* and *strong program* in cognitive science, where weak and strong refer to how strongly the two programs rely on the use of computational metaphors in describing cognition; for our purposes, these terms are analogous to constructivism and cognitivism.

6. The terms *theory-driven* and *data-driven* are being used loosely here and actually AI approaches should be seen as lying somewhere on a spectrum. Theory-driven approaches also use some kind of data (e.g., rule-based systems developed based on cognitive task analyses of human subjects), and some machine learning algorithms can be used in more theory-driven ways (e.g., how linear regression is typically used in statistics and the social sciences, including education research). Indeed, even strong proponents of rule-based and symbolic AI would not necessarily shy away from using theoretically motivated statistical techniques in their work; Newell and Rosenbloom's (1981) power law of practice is a prime example of this.

7. This tradeoff can also be seen in the context of learner modeling. The knowledge decomposition hypothesis coming from information-processing psychology (Anderson et al., 1998; Koedinger et al., 2013) has been highly influential, resulting in a variety of predictive models that are commonly used in intelligent tutoring systems and educational data mining research (e.g., Bayesian knowledge tracing, additive factors model, performance factors model). Indeed, these models resulted from the work of Anderson, Newell, and other cognitivists. Recently, researchers have shown that deep neural networks can provide a way to automatically discover complex interdependencies among knowledge components, using models such as deep knowledge tracing (Piech et al., 2015). However, such connectionist models have not been motivated by situativist or constructivist theories. Moreover, recent research has shown that while deep knowledge tracing can have higher predictive accuracy that other models, it does not actually detect meaningful interrelationships between skills (perhaps because of the high variance and limited amounts of data; Montero et al., 2018).

8. More precisely, an instructional intervention can be thought of as a policy in a Markov decision process, where at each time step, the instructor's action changes the state of the classroom, and the goal is to reach a state that results in a good educational experience for each student.

9. Note that constructivism is both a learning theory, as discussed above, and an epistemological position, due to the nature of the theory.

10. It was his own daughter.

11. One point of confusion might be that qualitative methods are known to have researcher "bias,'' since the researcher interacts with participants. Here such "bias" could possibly be thought of as variance, because researcher biases might vary from researcher to researcher. If several researchers with different backgrounds and agendas perform similar ethnographic studies, their biases might "cancel out" in a sense. This is not to deny that there may be systematic biases in qualitative research however; some qualitative research can certainly be quite biased and high variance. On the other hand, quantitative methods are said to be biased here, because they ignore the richness of the phenomena they are studying, a richness that qualitative methods can provide. The assumption here is that, at least in the views of qualitative researchers (like constructivists and situativists), such oversimplifications of the phenomena under study are more disconcerting than any systematic biases that may arise in qualitative research.

References

Aleven, V., & Koedinger, K. R. (2013). Knowledge component (kc) approaches to learner modeling. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 1. Learner modeling* (pp. 165–182). U.S. Army Research Laboratory.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18. https://doi.org/10.1037/a0021017

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher*, *25*(4), 5–11. https://doi.org/10.3102/0013189X025004005

Anderson, J. R., Reder, L. M., & Simon, H. A. (1997). Situative versus cognitive perspectives: Form versus substance. *Educational Researcher*, *26*(1), 18–21. https://doi.org/10.3102/0013189X026001018

Anderson, J. R., Reder, L. M., & Simon, H. A. (1998). Radical constructivism and cognitive psychology. *Brookings Papers on Education Policy*, *1998*(1), 227–278.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1999). *Applications and misapplications of cognitive psychology to mathematics education*. http://act-r.psy.cmu.edu/papers/misapplied.html

Bach, S. H., Broecheler, M., Huang, B., & Getoor, L. (2017). Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, *18*(109), 1–67. https://arxiv.org/abs/1505.04406

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1

Bernstein, R. (1983). *Beyond objectivism and relativism: Science, hermeneutics, and praxis*. University of Pennsylvania Press. https://doi.org/10.9783/9780812205503

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/BF00117832

Brockman, J. (1996). *Third culture: Beyond the scientific revolution*. Simon & Schuster. https://doi.org/10.1119/1.18425

Brown, A. L., & Campione, J. C. (1994). *Guided discovery in a community of learners*. MIT Press.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42. https://doi.org/10.3102/0013189X018001032

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Chan, D. (2017). The AI that has nothing to learn from humans. *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/10/alphago-zero-the-ai-that-taught-itself-go/543450/

Cobb, P. (1990). A constructivist perspective on information-processing theories of mathematical activity. *International Journal of Educational Research*, *14*(1), 67–92. https://doi.org/10.1016/0883-0355(90)90017-3

Cobb, P. (2007). Putting philosophy to work. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 2–38). Information Age.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Danish, J. A., & Gresalfi, M. (2018). Cognitive and sociocultural perspectives on learning: Tensions and synergy in the learning sciences. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 34–43). Routledge. https://doi.org/10.4324/9781315617572-4

Derry, S. J. (1996). Cognitive schema theory in the constructivist debate. *Educational Psychologist*, *31*(3–4), 163–174. https://doi.org/10.1207/s15326985ep3103&4_2

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2–3), 105–225. https://doi.org/10.1080/07370008.1985.9649008

Domingos, P., Kok, S., Poon, H., Richardson, M., & Singla, P. (2006). Unifying logical and statistical AI. In *Proceedings of the twenty-first AAAI conference on artificial intelligence* (pp. 2–7). AAAI Press. http://www.cs.washington.edu/homes/pedrod/papers/aaai06c.pdf

Donmoyer, R. (2006). Take my paradigm . . . please! The legacy of Kuhn's construct in educational research. *International Journal of Qualitative Studies in Education*, *19*(1), 11–34. https://doi.org/10.1080/09518390500450177

Doroudi, S., Holstein, K., & Johanes, P. (2020). Probing learning scientists' beliefs about learning and science. In M. Resalfi, & I. S. Horn (Eds.), *The interdisciplinarity of the learning sciences, 14th international conference of the learning sciences (ICLS) 2020* (*Vol. 1*, pp. 317–324). International Society of the Learning Sciences.

Dwivedi, R., Singh, C., Yu, B., & Wainwright, M. J. (2020). Revisiting complexity and the bias-variance tradeoff. https://arxiv.org/abs/2006.10189

Gagné, R. M., & Brown, L. T. (1961). Some factors in the programming of conceptual learning. *Journal of Experimental Psychology*, *62*(4), 313–321. https://doi.org/10.1037/h0049210

Geertz, C. (1973). *The interpretation of cultures*. Basic Books.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational Researcher*, *26*(1), 5–17. https://doi.org/10.3102/0013189X026001005

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*(1), 5–26. https://doi.org/10.1037/0003-066X.53.1.5

Greeno, J. G. (2015). Commentary: Some prospects for connecting concepts and methods of individual cognition and of situativity. *Educational Psychologist*, *50*(3), 248–251. https://doi.org/10.1080/00461520.2015.1077708

Hu, Z., Ma, X., Liu, Z., Hovy, E., & Xing, E. (2016). *Harnessing deep neural networks with logic rules*. https://arxiv.org/abs/1603.06318

Jacobson, M. J., Kapur, M., & Reimann, P. (2016). Conceptualizing debates in learning and educational research: Toward a complex systems conceptual framework of learning. *Educational Psychologist*, *51*(2), 210–218. https://doi.org/10.1080/00461520.2016.1166963

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Kersh, B. Y., & Wittrock, M. C. (1962). Learning by discovery: An interpretation of recent research. *Journal of Teacher Education*, *13*(4), 461–468. https://doi.org/10.1177/002248716201300417

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935–937. https://doi.org/10.1126/science.1238056

Kolodner, J. L. (2002). The "neat" and the "scruffy" in promoting learning from analogy: We need to pay attention to both. *Journal of the Learning Sciences*, *11*(1), 139–152. https://doi.org/10.1207/S15327809JLS1101_7

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, *43*(2), 393–427. https://doi.org/10.1017/S0022226707004628

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Lincoln, Y. S., Lynham, S. A., & Guba, E. G. (2017). Paradigmatic controversies, contradictions, and emerging confluences, revisited. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 108–150). Sage.

Mehta, J., & Fine, S. (2019). *In search of deeper learning: The quest to remake the American high school*. Harvard University Press. https://doi.org/10.4159/9780674239951

Minsky, M. (1991). Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine*, *12*(2), 34–34.

Minsky, M., & Papert, S. A. (1971). *Progress report on artificial intelligence*. https://web.media.mit.edu/~minsky/papers/PR1971.html

Montero, S., Arora, A., Kelly, S., Milne, B., & Mozer, M. (2018). *Does deep knowledge tracing model interactions among skills?* International Educational Data Mining Society.

Munter, C., Stein, M. K., & Smith, M. S. (2015). Dialogic and direct instruction: Two distinct models of mathematics instruction and the debate(s) surrounding them. *Teachers College Record*, *117*(11), 1–32.

Nathan, M. J., & Sawyer, R. K. (2014). Foundations of the learning sciences. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 21–43). Cambridge University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition (Carnegie Mellon Symposia on Cognition Series*, 1st ed., pp. 1–55). Lawrence Erlbaum.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice Hall.

Neyshabur, B., Tomioka, R., & Srebro, N. (2014). *In search of the real inductive bias: On the role of implicit regularization in deep learning*. https://arxiv.org/abs/1412.6614

Norvig, P. (2017). *On Chomsky and the two cultures of statistical learning (2011)*. http://norvig.com/chomsky.html

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books.

Papert, S. (1987a). Information technology and education: Computer criticism vs. technocentric thinking. *Educational Researcher*, *16*(1), 22–30. https://doi.org/10.3102/0013189X016001022

Papert, S. (1987b). Microworlds: Transforming education. In R. W. Lawler & M. Yazdani (Eds.), *Artificial intelligence and education: Vol. 1. Learning environments and tutoring systems* (pp. 79–94). Ablex.

Papert, S. (1988). A critique of technocentrism in thinking about the school of the future. In B. Sendov, & I. Stanchev (Eds.), *Children in the information age* (pp. 3–18). Elsevier.

Pea, R. D. (1987). The aims of software criticism: Reply to Professor Papert. *Educational Researcher*, *16*(5), 4–8. https://doi.org/10.3102/0013189X016005004

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 505–513). MIT Press.

Quartz, S. R. (1999). The constructivist brain. *Trends in Cognitive Sciences*, *3*(2), 48–57. https://doi.org/10.1016/S1364-6613(98)01270-4

Ray, W. E. (1961). Pupil discovery vs. direct instruction. *Journal of Experimental Education*, *29*(3), 271–280. https://doi.org/10.1080/00220973.1961.11010692

Salomon, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. *Educational Researcher*, *20*(6), 10–18. https://doi.org/10.3102/0013189X020006010

Schank, R. (1983). The current state of AI: One man's opinion. *AI Magazine*, *4*(1), 3.

Schank, R. (2016). Why learning sciences? In M. A. Evans, M. J. Packer, & R. K. Sawyer (Eds.), *Reflections on the learning sciences* (pp. 19–31). Cambridge University Press.

Schwartz, D. L., Lindgren, R., & Lewis, S. (2009). Constructivism in an age of non-constructivist assessments. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (pp. 46–73). Routledge.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, *27*(2), 4–13. https://doi.org/10.3102/0013189X027002004

Shaffer, D. W. (2017). *Quantitative ethnography*. Cathcart Press.

Shaffer, D. W., & Serlin, R. C. (2004). What good are statistics that don't generalize? *Educational Researcher*, *33*(9), 14–25. https://doi.org/10.3102/0013189X033009014

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, *20*(7), 2–16. https://doi.org/10.3102/0013189X020007002

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159. https://doi.org/10.1037/h0030806

Spiro, R. J., Coulson, R. L., Feltovich, P. J., & Anderson, D. K. (1988). *Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains* (Technical Report No. 441). Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Spiro, R. J., Feltovich, P. J., Feltovich, P. L., Jacobson, M. J., & Coulson, R. L. (1991). Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. *Educational Technology*, *31*(5), 24–33.

Taber, K. S. (2010). Constructivism and direct instruction as competing instructional paradigms: An essay review of Tobias and Duffy's *Constructivist Instruction: Success or Failure? Education Review*, *13*(8). https://edrev.asu.edu/index.php/ER/article/view/1418/89

Tobias, S., & Duffy, T. M. (2009). *Constructivist instruction: Success or failure?* Routledge.

Vera, A. H., & Simon, H. A. (1993). Situated action: A symbolic interpretation. *Cognitive Science*, *17*(1), 7–48. https://doi.org/10.1207/s15516709cog1701_2

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., . . .Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350–354. https://doi.org/10.1038/s41586-019-1724-z

von Glasersfeld, E. (1991). An exposition of constructivism: Why some like it radical. In G. E. Mobus (Ed.), *IFSR International series in systems science and systems engineering: Vol. 7. Facets of systems science* (pp. 229–238). Springer.

Winch, W. H. (1913). *Inductive versus deductive methods of teaching: An experimental research* (Educational psychology monographs, No. 11). Warwick & York.

Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial Intelligence*, *170*(18), 1256–1258. https://doi.org/10.1016/j.artint.2006.10.011

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Author

SHAYAN DOROUDI is an assistant professor at the University of California, Irvine School of Education and (by courtesy) Department of Informatics. He works at the intersection of educational data science, educational technology, and the learning sciences, and his work draws inspiration from the histories of these fields.