AERA Open
January-December 2024, Vol. 10, No. 1, pp. 1–21
DOI: 10.1177/23328584241303495
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2024. https://journals.sagepub.com/home/ero

Addressing Threats to Validity in Supervised Machine Learning: A Framework and Best Practices for Education Researchers

Kylie Anglin
University of Connecticut

Given the rapid adoption of machine learning methods by education researchers, and the growing acknowledgment of their inherent risks, there is an urgent need for tailored methodological guidance on how to improve and evaluate the validity of inferences drawn from these methods. Drawing on an integrative literature review and extending a well-known framework for theorizing validity in the social sciences, this article provides both an overview of threats to validity in supervised machine learning and plausible approaches for addressing such threats. It collates a list of current best practices, brings supervised learning challenges into a unified conceptual framework, and offers a straightforward reference guide on crucial validity considerations. Finally, it proposes a novel research protocol for researchers to use during project planning and for reviewers and scholars to use when evaluating the validity of supervised machine learning applications.

Keywords: *artificial intelligence, program evaluation, technology, validity/reliability, research methodology, machine learning, supervised learning, validity, measurement, artificial intelligence (AI), algorithmic bias*

Introduction

Education research is currently undergoing a transformation, with scholars taking advantage of powerful machine learning technologies to generate novel educational insights. Broadly speaking, these technologies involve the use of computers to identify patterns in data (Samuel, 1959). This includes both supervised learning, where the goal is to identify patterns in labeled data (e.g., graded essays) to predict the labels of new data (e.g., ungraded essays), and unsupervised learning, where the goal is to identify potentially unknown patterns in data without any preconceived labels (e.g., clustering essays based on their content). Both approaches offer opportunities for education researchers. Indeed, education literature featuring machine learning has increased exponentially in the past decade (Mcfarland et al., 2021), with themed special issues indicating enthusiasm among journal editors (Mcfarland et al., 2021; Reardon & Stuart, 2019) and major funding organizations, including the National Science Foundation and the Institute of Education Sciences, promoting this line of research via themed competitions (NCSER, 2021).

However, alongside the rush to explore machine learning's potential research benefits, there is an urgent need to evaluate the validity of inferences drawn from machine learning methods. There is a growing acknowledgment, for example, that supervised learning models, like their human counterparts, risk identifying particular patterns that promote stereotyping and the unfair distribution of resources (Kordzadeh & Ghasemaghaei, 2022; Van Giffen et al., 2022). For example, many educational outcomes from standardized test scores to college enrollment not only result from a student's motivation and intelligence but also from the quality of the educational opportunities provided to them—factors that are intricately related to race and socioeconomic status (Reardon, 2011). Thus, supervised models that are trained on real-world data reflecting these realities have the capacity to exacerbate existing biases (Suresh & Guttag, 2021). Further, algorithmic bias is not the only mechanism whereby machine learning applications might cause faulty inferences. There are myriad ways in which a researcher may err in drawing conclusions from these methods.

Yet, despite the rapid adoption of machine learning methods by education researchers and growing acknowledgment of the inherent risks of these methods, methodological guidance in the literature is limited. In part, this is perhaps because foundational papers in machine learning have been developed beyond the social sciences, rarely address education issues, and may not share education researchers' validity concerns. Further, although extensive guidance is available for education researchers as they plan and judge the quality of randomized experiments and quasi-experiments—from sources such as the What Works Clearinghouse (2019)—there is no such centralized guidance for researchers' use of machine learning.

Given these challenges, it makes sense to consider the standards by which our field ought to evaluate studies involving machine learning, studies that now cover topics across curriculum, pedagogy, and policy. A shared understanding of how to weigh these studies' claims and evidence can aid our interpretation of their scholarly contribution and the value of their recommendations for educational practitioners. Furthermore, methodological guidance that is tailored to education researchers' specific needs could improve the quality of machine learning-based studies in the first place.

This article offers a series of contributions toward these objectives, focusing specifically on supervised learning. The guiding framework underpinning this effort derives from the discussion by Shadish et al. (2002) of validity types and associated threats to validity. For decades, social scientists have relied on the validity-types framework to guide their thinking about valid impact estimates (Campbell & Stanley, 1963; Shadish et al., 2002). In this approach, researchers consider the validity of inferences in terms of (a) the constructs represented by variables (construct validity), (b) the strength of association between two variables (statistical validity), (c) the causal relationship of those variables (internal validity), and (d) the generalizability of that relationship (external validity).

This article builds on the validity-types framework by considering how these four types of inferences pertain to instances of supervised learning. For each type, we address the following questions:

- *Construct validity*: To what extent does a model reflect the construct it aims to predict? (Has the outcome of interest been defined and labeled appropriately? Do the model predictions align with this definition?)
- *Statistical validity*: What is a model's estimated performance, sensitivity, and uncertainty? (Are the performance metrics unbiased? How large is the sample on which performance was measured?)
- *External validity*: How generalizable is the model performance? (Can the model be applied in the necessary circumstances while retaining its predictive ability? Is the model's predictive ability consistent across subgroups?)
- *Internal validity*: To what extent are the discussed relationships between outcomes, predictors, and/or treatments causal? (Are there confounders of an observed correlation between the treatment and machine learning-based measures of the outcome? And, if interpreted as such, is the relationship between predictors and outcomes truly causal?)

Drawing on an integrative review of machine learning applications that have appeared in *American Education Research Association* (*AERA*) journals, this article discusses the implications of each validity type for supervised learning research—identifying important threats to validity and offering a list of approaches to protect against such threats. The article thus critically interprets emerging supervised learning challenges via a unified framework already familiar to education researchers. Finally, the article culminates in a research protocol that can be used by education researchers in the planning stages of a machine learning project as well as by reviewers and readers seeking to judge the validity of machine learning applications.

Theoretical Framework

Shadish et al. (2002, p. 34) defined *validity* as "the approximate truth of an inference." A foundational proposition of this article, therefore, is that the application of machine learning in education results in *inferences*—inferences about education, about education research, and about how these might be improved. Consider automatic grading systems, a common educational application of supervised learning. To develop such a system, researchers commonly ask human graders to rate a series of student essays according to their quality. These graded essays constitute the *gold-standard labeled data*, which the researchers anticipate their algorithm will learn to predict. The gold-standard data are randomly split into a *training* set and a *testing* set. Using the training data, the model learns a relationship between predictors (in this case, certain features of the written essays) and ratings. The correspondence between human and machine ratings is then assessed via the testing data, with researchers reporting the model's performance metrics (see, e.g., Valenti et al., 2003). Using such performance metrics, authors and readers then draw inferences about whether the algorithm can or should be used in educational practice.

Thus, using the validity-types framework, the validity of such an inference relies on construct validity (e.g., has "writing quality" been appropriately defined and labeled? To what extent do the automatic grader's predictions reflect the construct of "writing quality"?), external validity (e.g., in which populations and settings will the model's predictions be faulty?), and statistical validity (e.g., is the presented performance metric an unbiased estimate of model error? How much uncertainty surrounds that estimate?). If the automatic grader is later used to measure the impact of an intervention, internal validity is also required (e.g., does the correlation reflect a causal relationship?). In using the validity-types framework, a researcher considers each of these validity types in turn, probing and adjusting for corresponding threats.

Of course, the understanding of validity by Shadish et al. (2002) is one formulation among many and may not even be the most common conceptualization in education research. The *Standards for Educational and Psychological Testing* (Phelps, 2011), for example, draw from a conceptualization of validity that is closer to the scholarship of Kane (1992) and Messick (1989) and posit that validation is best understood "as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (Phelps, 2011, p. 11). In this line of thinking, researchers should (a) explicitly state the proposed interpretation of test scores, (b) identify the inferences and assumptions required to make a leap from the scores to the interpretation, (c) assemble all available evidence relevant to the inferences and assumptions, (d) evaluate the most problematic assumptions in the argument, and (e) continue to adjust the argument or interpretation as necessary (Kane, 2001).

When fully implemented in measurement scenarios, this alternative approach to construct validity is more comprehensive than the validity-types framework. Although researchers using the validity-types framework would only consider the listed threats of Shadish et al. (2002), a researcher successfully implementing an argument-based validation approach would consider all necessary assumptions, focusing on the most relevant ones to the test's proposed use. However, Kane and Mesick's conceptualization of validity speaks less to inferences other than those drawn from scores on tests (e.g., omitting inferences about causality). Further, creating a comprehensive validity argument is not straightforward (Kane, 1992, 2001). Thus, although an argument-based approach to validation may be more comprehensive and theoretically ideal in some scenarios, the checklist-like approach to validation by Shadish et al. (2002)—where researchers consider each threat in turn, checking off those that they have ruled out—is a more practical heuristic for our purposes. Thus, the protocol presented in the final section of this article provides such a checklist with specific questions to consider related to each validity type when planning and evaluating a study using supervised learning.

Approach and Organization

In this article, the validity-types framework is used to organize and contextualize threats to validity in educational applications of supervised learning. The threats-to-validity discussion draws on the framework and an integrative, restricted review of supervised learning applications of Shadish et al. (2002) in academic journals published by AERA—the largest American professional society focused on education research (AERA, 2024). The review includes studies published in the *American Educational Research Journal, Educational Researcher, Educational Evaluation and Policy Analysis, Journal of Educational Behavior and Statistics*, and *AERA Open*. Figure 1 provides an overview of the search and exclusion parameters. The final set of studies is limited to 27 articles, which either trained or used a supervised learning model to answer an education research question via the analysis of nonsimulated educational data. An additional 11 methodological and/or conceptual articles were consulted and cited where relevant. A full list of reviewed articles can be found in Tables 1 and 2. It is important to note that the review is not intended as a meta-analysis, nor is it meant to test a theory or formally summarize the state of the literature. Instead, the studies are used to illustrate threats to validity and current best practices for addressing those threats.

The following sections discuss each validity type in turn—construct, external, statistical, and internal—within a supervised learning context. Each of these sections describes threats to validity and outlines common methodological approaches to addressing those threats. A summary of the validity types, alongside illustrative examples, can be found in Table 3. Then, drawing on the identified threats and best practices, the article concludes with a presentation of a research protocol: a series of questions for researchers and reviewers to consider when conducting and evaluating supervised learning applications in education.

Construct Validity in Supervised Learning

Often, when researchers apply supervised learning in educational contexts, it is for a measurement purpose: Researchers have a specific construct they aim to measure and train a supervised learning algorithm to do so (e.g., researchers might use an automated essay grader to measure essay quality). The validity of the resulting conclusions thus relies on the *construct validity* of the resulting supervised learning measure. Shadish et al. (2002, p. 20) defined *construct validity* as the validity of "inferences about the constructs that research operations represent." For example, beyond supervised learning applications, researchers commonly operationalize "teaching quality" using teaching observation rubrics. In such cases, construct validity concerns the extent to which the observation rubric truly reflects the construct of interest (teaching quality).

In studies involving supervised learning, there is often a secondary level of operationalization. Researchers begin with an initial operationalization of a construct using traditional means, and then they use a machine learning algorithm to replicate those measures. For example, researchers may use observation rubrics to operationalize teaching quality and then train an algorithm to replicate those observation scores. To make valid inferences regarding teaching quality in these cases, we must infer that (a) the observation scores appropriately capture teaching quality and (b) the supervised learning algorithm has retained the prototypical features of teaching quality that were captured by the observation scores. Threats to construct validity may occur at either stage—from the construct to measure or from the measure to supervised learning prediction. Four original threats from Shadish et al. (2002) therefore remain relevant: the inadequate explication of constructs, confounding constructs, mono-operation and monomethod bias, and participant

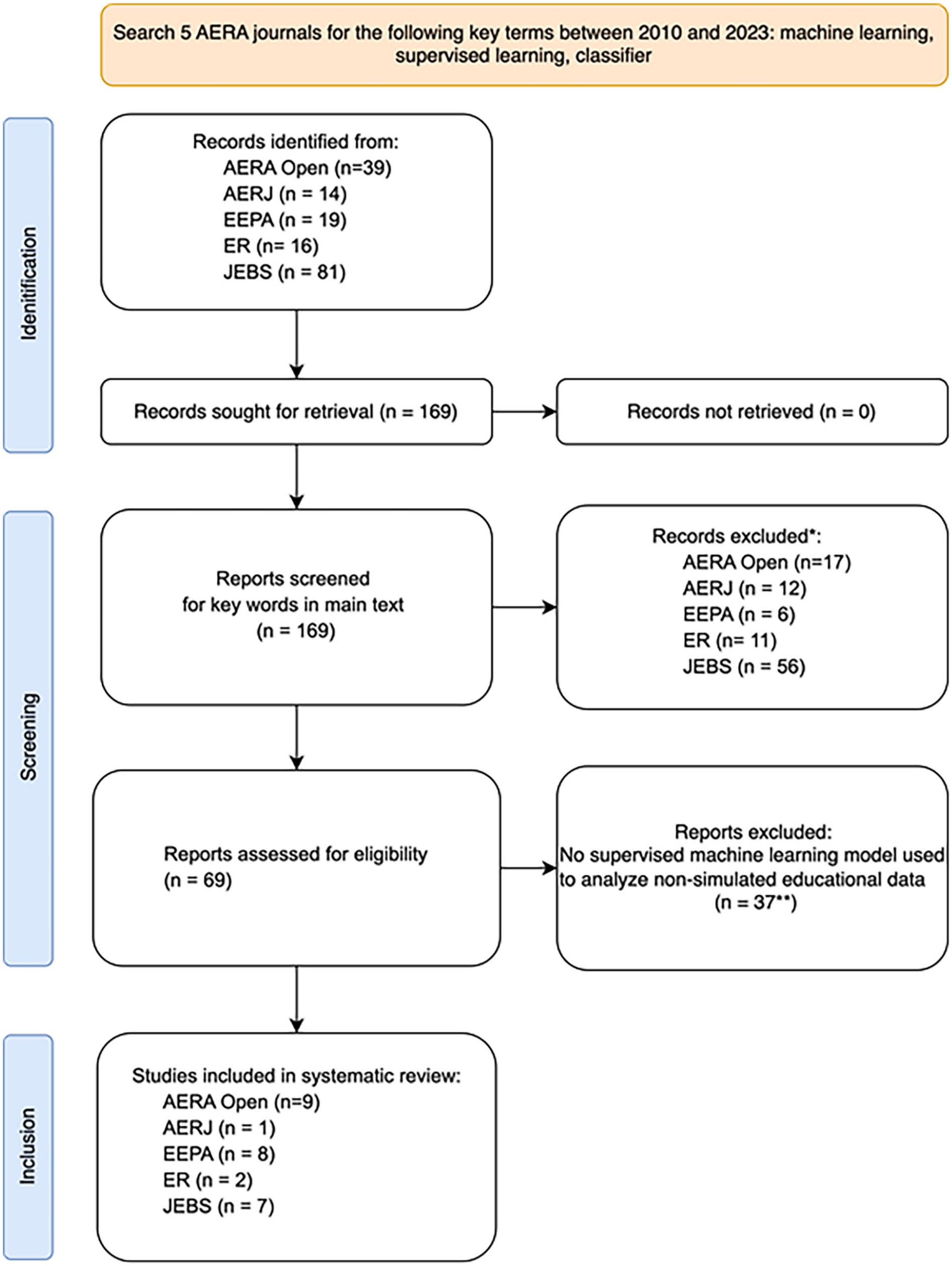
Search 5 AERA journals for the following key terms between 2010 and 2023: machine learning, supervised learning, classifier
Identification
Records identified from:
AERA Open (n=39)
AERJ (n = 14)
EEPA (n = 19)
ER (n= 16)
JEBS (n = 81)
Records sought for retrieval (n = 169)
Records not retrieved (n = 0)
Screening
Reports screened for key words in main text (n = 169)
Records excluded*:
AERA Open (n=17)
AERJ (n = 12)
EEPA (n = 6)
ER (n= 11)
JEBS (n = 56)
Reports assessed for eligibility (n = 69)
Reports excluded:
No supervised machine learning model used to analyze non-simulated educational data
(n = 37**)
Inclusion
Studies included in systematic review:
AERA Open (n=9)
AERJ (n = 1)
EEPA (n = 8)
ER (n = 2)
JEBS (n = 7)

FIGURE 1. *Search Protocol and Inclusion Criteria for Integrative Review.*

[a] Records were excluded if no keywords were found in the main text, excluding bibliographies and author biographies.

[b] Count includes unsupervised learning applications (n = 8) and conceptual or methodological articles focused on machine learning but lacking nonsimulated education data (n = 11).

Source: Figure adapted from PRISMA diagram (Page et al., 2021, p. 5).

TABLE 1
Applied Articles Reviewed

Journal	Authors	Title	Year
AERA Open	González Canché, M. S.	The geography of mathematical (dis)advantage: An application of multilevel simultaneous autoregressive (MSAR) models to public data in education research.	2023
AERA Open	Gormley, W. T., Jr., Amadon, S., Magnuson, K., Claessens, A., & Hummel-Price, D.	Universal pre-K and college enrollment: Is there a link?	2023
AERA Open	Lang, D., Wang, A., Dalal, N., Paepcke, A., & Stevens, M. L.	Forecasting undergraduate majors: A natural language approach.	2022
AERA Open	Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y.	Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education.	2021
AERA Open	Rosenberg, J. M., Borchers, C., Dyer, E. B., Anderson, D., & Fischer, C.	Understanding public sentiment about educational reforms: The Next Generation Science Standards on Twitter.	2021
AERA Open	Lucy, L., Demszky, D., Bromley, P., & Jurafsky, D.	Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas US history textbooks.	2020
AERA Open	Cimpian, J. R., & Timmer, J. D.	Large-scale estimates of LGBQ–heterosexual disparities in the presence of potentially mischievous responders: A preregistered replication and comparison of methods.	2019
AERA Open	Ramirez, G., Hooper, S. Y., Kersting, N. B., Ferguson, R., & Yeager, D.	Teacher math anxiety relates to adolescent students' math achievement.	2018
AERA Open	Page, L. C., & Gehlbach, H.	How an artificially intelligent virtual assistant helps students navigate the road to college.	2017
AERJ	Chen, D., Hebert, M., & Wilson, J.	Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application.	2022
EEPA	Jabbari, J., Chun, Y., Huang, W., & Roll, S.	Disaggregating the effects of STEM education and apprenticeships on economic mobility: Evidence from the LaunchCode program.	2023[a]
EEPA	Pietsch, M., Aydin, B., & Gümüş, S.	Putting the instructional leadership–student achievement relation in context: A meta-analytical big data study across cultures and time.	2023[a]
EEPA	Anglin, K.	The role of state education regulation: Evidence from the Texas Districts of Innovation statute.	2023
EEPA	Chi, O. L., & Lenard, M. A.	Can a commercial screening tool help select better teachers?	2023
EEPA	Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C.	Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course.	2024
EEPA	Master, B. K., Schwartz, H., Unlu, F., Schweig, J., Mariano, L. T., Coe, J., Wang, E. L., Phillips, B., & Berglund, T.	Developing school leaders: Findings from a randomized control trial study of the Executive Development Program and paired coaching.	2022
EEPA	Lee, J. C., Dell, M., González Canché, M. S., Monday, A., & Klafehn, A.	The hidden costs of corroboration: Estimating the effects of financial aid verification on college enrollment.	2021
EEPA	Liu, J., & Cohen, J.	Measuring teaching practices at scale: A novel application of text-as-data methods.	2021
ER	Kelly, S., & Abruzzo, E.	Using lesson-specific teacher reports of student engagement to investigate innovations in curriculum and instruction.	2021
ER	Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K.	Automatically measuring question authenticity in real-world classrooms.	2018
JEBS	Arthur, D., & Chang, H.-H.	DINA-BAG: A bagging algorithm for DINA model parameter estimation in small samples.	2024[a]
JEBS	Mozer, R., Miratrix, L., Relyea, J. E., & Kim, J. S.	Combining human and automated scoring methods in experimental assessments of writing: A case study tutorial.	2023
JEBS	Si, Y., Little, R. J., Mo, Y., & Sedransk, N.	A case study of nonresponse bias analysis in educational assessment surveys.	2023
JEBS	Suk, Y., Kim, J.-S., & Kang, H.	Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes.	2021
JEBS	Wu, E., & Gagnon-Bartsch, J. A.	Design-based covariate adjustments in paired experiments.	2021
JEBS	Sales, A. C., Hansen, B. B., & Rowan, B.	Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates.	2018
JEBS	Strobl, C., Wickelmaier, F., & Zeileis, A.	Accounting for individual differences in Bradley-Terry models by means of recursive partitioning.	2011

[a] Indicates Online First at time of search.

TABLE 2
Conceptual/Methodological Articles Reviewed

Journal	Authors	Title	Year
AERA Open	McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A.	Education data science: Past, present, future.	2021
AERA Open	Doroudi, S.	The bias–variance tradeoff: How data science can inform educational debates.	2020
AERA Open	Cope, B., & Kalantzis, M.	Big data comes to school: Implications for learning, assessment, and research.	2016
JEBS	Rothacher, Y., & Strobl, C.	Identifying informative predictor variables with random forests.	2024[a]
JEBS	Suk, Y., & Han, K. T.	A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning.	2024[a]
JEBS	Doran, H.	A collection of numerical recipes useful for building scalable psychometric applications.	2023
JEBS	Li, X., Xu, H., Zhang, J., & Chang, H.	Deep reinforcement learning for adaptive learning systems.	2023
JEBS	Pang, B., Nijkamp, E., & Wu, Y. N.	Deep learning with TensorFlow: A review.	2020
JEBS	Hao, J., & Ho, T. K.	Machine learning made easy: A review of Scikit-learn package in Python programming language.	2019
JEBS	Von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H.	Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities.	2019
JEBS	Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D.	Tools for educational data mining: A review.	2017

[a]Indicates online first at time of search.

reactivity. A related threat in supervised learning is also worth being made explicit here: When there are errors in the gold-standard data, there will necessarily be errors in the final supervised learning measure. These threats are discussed next.

Inadequate Explication of Constructs

Measurement scholars have long acknowledged that valid measurement is bolstered by a strong theoretical understanding of the construct being studied (Cronbach & Meehl, 1955). Thus, a foundational step for improving construct validity in any measurement exercise is the careful specification of the theoretical construct of interest. Shadish et al. (2002, p. 74) considered a failure to do so as the "inadequate explication of constructs." Given that the first level of operationalization in a supervised learning application involves turning a theoretical construct into labels within the gold-standard (training/testing) data, carefully specifying the construct of interest allows researchers to improve the quality of the gold-standard data and allows readers to assess the quality of model output. Researchers take two common approaches to addressing this threat:

- They can provide a comprehensive definition of the construct of interest. For example, when using supervised learning to measure "authentic questioning," Kelly et al. (2018, p. 452) defined *authentic questioning*—within the context of dialogic instruction—as "questions for which the answers are not presupposed by the teacher," and they linked this definition to several instructional frameworks for effective teaching, thereby identifying the literature to which their study spoke.
- They can acknowledge any debate or challenges in operationalizing the construct. For example, in predicting graduation, Bird et al. (2021, p. 3) explained the difficulty of defining "drop-out" given that students often leave college for periods of time while intending to return. Thus, the researchers instead aimed to predict "graduation," where *graduation* was defined as completing "any college-level credential within 6 years" [Bird et al., 2021, p. 3], and they also provided an evidence-based justification for this definition, drawing on national time-to-completion data.

Errors in Human Labels

In the social sciences, gold-standard data are often created by researchers via hand labeling according to the construct of interest. In the qualitative literature, the process of applying labels to data is typically referred to as *coding*

TABLE 3
Summary of Validity Types.

Validity type	Definition	Example from literature
Construct validity	Validity of inferences regarding the extent to which a model reflects the construct it is aimed at predicting	Kelly et al. (2018) developed a machine learning-based measure of *authentic questioning* (the construct of interest). The construct validity of this measure depends on the extent to which (a) their gold-standard human labels of authentic questions are aligned with the provided definition of the construct and (b) the machine learning algorithm has retained the prototypical features of authentic questioning. Steps they took to address construct validity included providing the reader with a definition and example of authentic questioning, linking their construct definition to the wider literature, training human coders to use the provided codebook, reporting measures of agreement between coders, and limiting the machine learning model to theoretically relevant predictors.
External validity	Validity of inferences regarding the generalizability of model performance	Liu and Cohen (2021) aimed to develop generalizable automated measures of effective teaching, including training a supervised learning model to identify open-ended questions. The external validity of this model depends on the extent to which the predictive validity of the model generalizes beyond the training data—to the population of teachers for whom they hoped the model would be useful. Steps they took to address external validity included maximizing alignment between the sample and target population, describing the source of their training and testing data (including the representation of important subgroups), and testing the performance of their model on hold-out data.
Statistical validity	Validity of inferences regarding the estimated performance, sensitivity, and uncertainty surrounding a model	Bird et al. (2021) trained a classifier aimed at predicting graduation and used an independent testing dataset to estimate the performance of the model. The statistical validity of their study depends on the valid estimation of model error and an appropriate understanding of the degree of confidence that those estimates warrant. Steps they took to address statistical validity included presenting multiple performance metrics (accuracy, precision, recall, and F_1 score, among others), using a large testing dataset of $>$33,000 students, and assessing the sensitivity of inferences to model parameters.
Internal validity	Validity of inferences regarding causal relationships between predictors, treatments, and outcomes	Master et al. (2022) trained a causal forest to identify heterogeneous effects of a principal professional development program. If the aim of this analysis was theory generation, then the internal validity of findings regarding potential moderators depended on whether the identified predictors actually produced the observed heterogeneity. The authors were careful not to overstate causal claims with their findings but took several steps to address instability in predictor importance (increasing readers' confidence that the authors had identified the most important measured moderators). These steps included using an ensemble model and testing the predictive ability of the identified characteristics in a hold-out sample.

(while *labeling* or *annotation* are more commonly used in machine learning (Anglin et al., 2022). Although often overlooked in the machine learning literature, where fallible human labels may be treated as "ground truth" (Geiger et al., 2020, p. 325; Zheng et al., 2024), the coding process is central to determining the validity of supervised learning predictions. At best, a supervised learning algorithm can only learn to replicate human codes. However, as decades of qualitative research have demonstrated, human coding is rarely a straightforward process because codes are contextual, theoretical, and contestable (Shaffer & Ruis, 2021). Many rigorous qualitative research practices are thus also applicable here. Researchers can do the following:

- They can provide a comprehensive codebook for human labeling (as in Aulck et al., 2021; Kelly et al., 2018; Nystrand et al., 1997). A *codebook* is a set of coding instructions that provides a definition of each label alongside examples and nonexamples (Shaffer & Ruis, 2021). For example, Kelly et al. (2018) created a codebook for labeling authentic questions that was 74 pages long and provided specific instructions to coders about how to handle common ambiguous teacher questions such as "What else?" (see Nystrand [2004] and Nystrand et al. [1997] for details on the codebook).
- They can disclose measures of agreement between multiple human labelers (as undertaken by Kelly et al., 2018; Liu & Cohen, 2021; Ramirez et al., 2018). A high level of agreement indicates that multiple labelers' understandings of the construct's definition are closely aligned (Shaffer & Ruis, 2021).

Relevant metrics include simple agreement, Krippendorf's alpha, Cohen's kappa, and correlation coefficients (Krippendorff, 2004).
- They can describe human labelers' training, knowledge, perspectives, and experience, allowing readers to gauge whether the labelers have the necessary knowledge and experience to understand a construct (Shaffer & Ruis, 2021; Snow et al., 2008).

Confounding Constructs

Confounding is typically understood in the context of internal validity, occurring when the correlation between a presumed cause (Variable A) and presumed effect (Variable B) is due to a third variable that is correlated with Variables A and B. The presumed causal relationship, then, is confounded by the extraneous variable. Shadish et al. (2002) argued, however, that the interpretation of constructs also may be confounded by extraneous variables. They provided the example of describing a sample as "unemployed"; the sample may indeed be limited primarily to those without jobs but also may disproportionately include victims of racial prejudice. Interventions that aim to address only one aspect of unemployment (e.g., currently jobless) are likely to be of limited use if the other construct (e.g., discrimination) proves to be a greater determinant. In this case, a construct validity error would occur if only one of the constructs is acknowledged.

In supervised learning applications, when construct confounding occurs at the first level of operationalization (from construct to measure), confoundedness may be exacerbated at the second level of operationalization (from measure to machine learning prediction). Consider, for example, the challenge of predicting college graduation. In most colleges and universities, drop-out occurs more frequently among Black, Hispanic, and lower-income students (Bird et al., 2021). Thus, as with the preceding unemployment example, drop-out is confounded by demographic characteristics. If demographic characteristics are included in the model, the model likely would identify these demographic variables as key predictors, resulting in students of color being more likely to be labeled as at risk for dropping out regardless of whether other associated risk factors are present (Baker & Hawn, 2021). Further, even if a researcher excludes demographic variables from the model, the model may focus on theoretically irrelevant factors that correlate with demographic variables (Hovy & Spruit, 2016). This phenomenon is one of the most commonly discussed types of algorithmic biases in the machine learning literature, variously termed *social bias, historical bias, societal bias*, or *preexisting bias* (Van Giffen et al., 2022).

It is worth briefly considering, however, why and when construct confounding is a problem for construct validity rather than, say, an instance of effective prediction. After all, the aim of supervised learning is to predict an outcome by identifying existing patterns. In the example, the model isn't wrong to predict that students of color are more likely to drop out; in fact, because of systemic factors, they are (Brown & Rodríguez, 2009). The validity error would come in the interpretation of the label, particularly in the researcher's failure to acknowledge the relationship between race, socioeconomic status (SES), and schooling (Bradley & Renzulli, 2011) despite machine learning predictions for individuals being influenced by these factors.

Importantly, construct confounding also can result from idiosyncrasies in the creation of training data, even in the absence of any real-world co-occurrence of constructs. Consider one infamous example. In "Automated inference on criminality using face images," researchers claimed successful use of supervised learning to draw inferences about the criminality of individuals from photographs of their faces (X. Wu & Zhang, 2016, p. 10). However, critics later pointed out that the noncriminal photographs were selected from personal and professional websites, where people are commonly smiling, whereas the criminal photographs were selected from formal identification sources (e.g., driver's license photos), where smiling was less common (Bergstrom & West, 2021; Bowyer et al., 2020). In other words, in the training data, "criminality" was confounded by smiling (even though smiling may not necessarily correlate with criminality outside these data); it was smiles, not criminality, that the classifier could identify. Concluding that a classifier can identify "criminality" rather than smiling is therefore erroneous, as is the conclusion that "it is possible to infer character from features" (X. Wu & Zhang, 2016, p. 1).

To address the threat of confounding constructs, researchers can do the following:

- They can limit predictors to those that are theoretically relevant. For example, in predicting authentic questioning, Kelly et al. (2018, p. 455) limited themselves to "theoretically grounded language features" such as question stems and parts of speech tags. A supervised learning measure is less likely to be confounded by an extraneous nuisance variable if the researcher restricts the model to factors that are theoretically relevant to the construct (Zheng et al., 2024).
- They can assess predictor importance using interpretable algorithms. For example, Lang et al. (2022) used data ablation techniques, systematically varying the predictors incorporated in their college major classifier to determine which predictors were most important. If a predictor without theoretical relevance to the outcome surfaces, this may indicate a co-occurring and potentially misleading construct (see also Bowyer et al., 2020; X. Wu & Zhang, 2016).

- They can assess the fairness of the model using formal approaches, including statistical parity, separation, and differential algorithmic functioning (Barocas et al., 2023; Suk & Han, 2024).

Mono-operation and Monomethod Bias

All measures underrepresent constructs and contain irrelevancies (Shadish et al., 2002). For this reason, researchers are advised to use several measures of a given construct. Shadish et al. (2002) conceptualized a failure to do this as *mono-operation bias* (relying on a single measure) associated with *monomethod bias* (relying on a single method of measurement). For example, readers of a study may be suspicious if an intervention improves a construct when that construct is only measured using self-report. A stronger approach may be to triangulate results from both self-report and teacher report. The same advice holds true when supervised learning is used to measure an outcome. Construct validity will increase when there are multiple measures and methods of measurement, especially where these span both human and machine approaches (Grimmer & Stewart, 2013). To address mono-operation and monomethod bias, researchers commonly do the following:

- They can replicate findings obtained with machine learning measures using non–machine learning-based measures (as in Mozer et al., 2023; Shores & Steinberg, 2022). For example, in estimating the number of student-weeks spent in remote instruction during the COVID-19 pandemic, Shores and Steinberg (2022) triangulated text classification–based estimates (applied to school websites) with mobile phone data—with key research findings consistent across both sources.
- They can probe the sensitivity of individual predictions to multiple algorithms. For example, in the work of Bird et al. (2021) on graduation prediction, the authors assessed the extent to which the relative ranking of students' drop-out risk was consistent across algorithms. Instability here would indicate that a decision of whether to intervene with a given student—because they are in the top *x* percentile for predicted drop-out risk, for example—may depend on the specific algorithm employed by the college.

Reactivity to the Machine Learning Model

Because humans actively interpret their surroundings and adapt their behavior in response, Shadish et al. (2002, p. 73) cautioned that "participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation"—a phenomenon known as *participant reactivity*. For example, psychological evaluation may cause participants to act or answer questions in ways they hope will be viewed as psychologically healthy (Rosenberg, 1969). A similar phenomenon can occur when participants learn that they are being evaluated by a machine learning model; participants may attempt to game the model by guessing the actions that will improve their score. For example, in some automated grading systems, longer essays often receive higher scores (Bridgeman et al., 2012). If this becomes common knowledge, participants may start writing longer essays without changing the underlying quality of the work (Cope & Kalantzis, 2016). To address this challenge, researchers can do the following:

- They can avoid sharing information about the method of measurement with participants. For example, when measuring the relationship between authentic questioning and teacher-reported student engagement, Kelly and Abruzzo (2021, p. 311) ensured that "teachers had no knowledge of the measures of instruction at the time of reporting." If participants are unaware of assessment specifics, they are less likely to successfully manipulate their scores. In contrast, when institutions use algorithms for high-stakes decision making, publicizing information on the predictors is also an important aspect of transparency and accountability (Zheng et al., 2024).
- They can aim for theoretical alignment between predictors and the construct of interest (as in Kelly et al., 2018). Given the conflict between transparency and participant reactivity, a better approach may be to ensure alignment between the predictors and the construct. In this way, reactivity can be directed toward more productive ends.
- They can conduct interviews and surveys with participants. It is impossible to prevent respondents from generating their own hypotheses regarding researcher intentions and from changing their behavior accordingly. However, reactivity may at least be probed through interviews or surveys of participants (Shadish et al., 2002).

External Validity in Supervised Learning

Shadish et al. (2002, p. 83) conceptualized external validity within a causal evaluation framework, defining it as the "extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes." In supervised learning, however, it is not a *causal* relationship that must hold over relevant variations but a *predictive* relationship. External validity in supervised learning thus may be conceptualized as the extent to which a model's predictive ability, as estimated using the provided performance metrics, generalizes to the intended use cases. In a single study, this means that the performance metrics—estimated using the labeled testing data—must be a good estimate of the performance of the model in the unlabeled data. In other words, model performance must generalize from the testing data to the full sample of data included in the study (Yarkoni & Westfall, 2017). These data may include variations in people, settings, and time (Kapoor & Narayanan, 2023). Further, when models are made available for public use and/or applied to novel datasets, the scenarios may become increasingly diverse.

Threats to external validity are the reasons that such generalizations may fail. Shadish et al. (2002) identified *interactions*—when three or more variables influence each other—as the key challenge to external validity. In a randomized experiment, external validity is threatened if there is a substantial coefficient on a three-way interaction between the treatment, the outcome, and a certain characteristic of either the unit, treatment, outcome, or setting. In supervised learning, the external validity threat similarly occurs when there is three-way interaction between the predicted outcome, predictors, and the characteristics of samples, settings, or time points. Just as treatment effects often vary with study characteristics (Bloom & Michalopoulos, 2013), so too do predictive relationships (Kapoor & Narayanan, 2023). Thus, three types of interaction effects—samples, settings, and time—are discussed in more detail next. One additional threat, particular to supervised learning applications, is also discussed: the failure of a model to generalize because it was overfit to noise in the training sample.

Interaction Between the Predictive Relationship and Variations in the Sample

In supervised learning, the process of estimating performance metrics implicitly assumes that the testing data are a random sample of the population to which the algorithm will be applied (Zadrozny, 2004). Yet, in many supervised learning applications, training and testing data are not a random sample of the population of interest and instead may have distinct characteristics—a phenomenon known as *sample selection bias*. When these characteristics moderate the relationship between predictors and the predicted outcome, external validity is threatened. Further, external validity requires that relationships generalize not only to new relevant populations but also to variations *within* the original population (Shadish et al., 2002). In other words, external validity is also threatened when the model exhibits differential performance for one or more represented subgroups. To address this threat, researchers commonly can do the following:

- They can select training/testing data so as to maximize alignment with the target population. Then, they can describe the source and characteristics of these data. For example, in developing automated approaches to measuring effective teaching, Liu and Cohen (2021) described the demographic characteristics of both the teachers and the students in their sample. They also noted the limitations of their classroom sample—fourth and fifth grade English language arts classrooms—indicating that “classroom discourse may well look different in mathematics or in the primary grades” (Liu & Cohen, 2021, p. 606).
- They can ensure sufficient representation among population subgroups. If there is an interaction between the predictive relationships within a model and model subgroups, the model must be provided with enough data to learn those interactions (Buolamwini & Gebru, 2018). This can be addressed by oversampling important subgroups. In the case of Liu and Cohen (2021), for example, ensuring the model’s generalizability across linguistic subpopulations might mean oversampling classrooms with high proportions of English language learners.
- They can evaluate the performance of the algorithm among subgroups (as in Chen et al., 2022; Lang et al., 2022). In addition to presenting average performance metrics, best practice requires that researchers also present performance metrics within subgroups (Mitchell et al., 2019). For example, Chen et al. (2022) assessed the performance of an automated essay scoring system among struggling and nonstruggling writers and demonstrated that the model was less reliable when scoring the essays of struggling writers. In other cases, additional subgroups might include those defined by race/ethnicity, nationality, gender, SES, and/or disability (Baker & Hawn, 2021).

Interaction Between the Predictive Relationship and Variations in Setting

Machine learning researchers often transport models trained in one setting for use in another (Lucy et al., 2020). For example, researchers commonly apply pretrained sentiment models to new data, such as applying a model trained to identify positive versus negative Yelp reviews to assess positive versus negative sentiment in student surveys. However, the sentiment of a particular word is often context dependent, creating an interaction between the setting and predictive relationships in a sentiment model. To address this threat, researchers can do the following:

- They can set aside a hold-out setting for validation. For example, Kelly et al. (2018) trained their authentic question classifier on one set of schools and validated the model on a hold-out school not used to train the classifier. If the model performs well in the hold-out setting, this indicates that predictors of authentic questioning can generalize across setting characteristics.

- They can evaluate pretrained models in the current setting. This may require hand labeling a sample of the current data to examine the performance of a pretrained classifier. For example, Lucy et al. (2020) evaluated a pretrained named-entity-recognition classifier (designed to identify proper nouns) to assess its ability to identify the names of people within history textbooks, finding that the performance was meaningfully lower than the performance on the original testing sample.

Interaction Between the Predictive Relationship and Variations in Timing

In many supervised learning applications, a model is trained on past data with the intention of applying it to future data. However, models that perform well initially may not retain their performance over time (Sculley et al., 2014), a phenomenon known as *drift* (Gama et al., 2004). A canonical example of model drift is the failure of Google Flu Trends. At one point, this model could accurately predict Centers for Disease Control and Prevention flu prevalence estimates days ahead of the release of the estimates (Ginsberg et al., 2009). Later, however, the model massively overestimated flu prevalence. The reasons for the failure of Google Flu Trends are not known, but one hypothesis is that changes in Google's search platform—for example, the incorporation of suggested search terms for users—dramatically changed the nature of the underlying search data (Lazer et al., 2014). As a result, the relationship between the predictors (search terms) and the predicted outcome (flu prevalence) proved unstable over time. In education contexts, policy changes might similarly influence the relationship between predictors and outcomes. For example, high-quality teaching might look and sound different following the adoption of Common Core standards (Cohen et al., 2022). In this situation, a supervised learning model trained in the pre–Common Core era may not perform well in the post–Common Core period. To address this threat, researchers commonly can do the following:

- They can assess the correlation between model performance and time (as in Lang et al., 2022). To evaluate the plausibility of model drift, researchers can assess whether there is a substantial correlation between model performance and time in past data. If the predictive ability holds stable over time in past data, this provides evidence that the predictive ability will be stable in future data.
- They can monitor model performance. Just as researchers should evaluate model performance in new settings, they should periodically evaluate model performance in new time periods (Sculley et al., 2014).
- They can update model training with current data. If model performance deteriorates, researchers can either retrain the model or update past training data with newly collected data (Lwakatare et al., 2020).

Model Overfit

Finally, all generalizations will be invalid if the model is overfit to the training data. When flexible algorithms are trained on data with many variables, an algorithm can reduce error in the training sample by learning idiosyncratic and ungeneralizable patterns (Hastie et al., 2009). Indeed, using its training data, a sufficiently flexible model can reduce error to zero without necessarily identifying any generalizable patterns. This is why a minimum standard for rigorous supervised learning incorporates the training/testing split. When model performance is estimated on data that are independent from the training data, these performance metrics provide a more accurate estimate of model generalizability (Emmert-Streib & Dehmer, 2019; Yarkoni & Westfall, 2017).

Statistical Conclusion Validity in Supervised Learning

In quantitative social science research, conclusions are drawn from statistical estimation, including point estimates (e.g., effect sizes), measures of uncertainty (e.g., standard errors), and statistical tests (e.g., null-hypothesis statistical testing). Statistical conclusion validity concerns the appropriateness of conclusions drawn from such evidence. In supervised learning, conclusions are similarly drawn from statistical estimation. Most commonly, conclusions regarding a model's usefulness are based on the magnitude of performance metrics (e.g., accuracy, precision, recall, etc.). Threats to statistical validity in supervised learning include situations in which we may over- or underestimate the magnitude of the performance metric or the degree of confidence that the performance metric warrants. Importantly, an incorrect understanding of performance can result in faulty decisions, including the deployment of a deficient model because its performance was overestimated or because confidence was overstated (Varoquaux, 2018). Four threats to statistical validity in supervised learning are discussed next: misleading or uninformative performance metrics, optimizing a model to the testing data, dependence between the training and testing data, and an insufficient testing data sample size.

Misleading or Uninformative Performance Metrics

Researchers can choose several performance metrics to gauge a model's usefulness. With binary classifiers—for example, classifying a student as at risk/not at risk—performance metrics commonly concern the relationship between true positives (TPs; positive cases correctly classified as positive according to the gold-standard data), true negatives

(TNs; negative cases correctly classified as negative), false positives (FPs; negative cases incorrectly classified as positive), and false negatives (FNs; positive cases incorrectly classified as negative). Metrics include

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

$$\text{Recall / sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FP rate} = 1 - \text{specificity}$$

(Note, however, that use of the term *true* here is somewhat misleading, a high rate of TNs and TPs indicates only agreement with the gold-standard data, which itself may be flawed, as discussed previously under construct validity.)

Researchers also commonly calculate summary statistics, such as the F_1 statistic and the area under the receiver operating characteristic curve (Manning & Schütze, 1999). When a supervised learning algorithm aims to predict a continuous outcome—for example, predicting a student's score for a given essay response—common performance metrics include the raw error, mean squared error, root mean squared error, and R^2.

Conclusions regarding a model's usefulness depend on an accurate understanding of the prevalence, magnitude, and types of error involved. How often, for example, does the model fail to identify an at-risk student? How distant is the average predicted teacher observation score from the gold-standard human observation score? Validity is threatened when the presented performance metrics omit this information. For example, accuracy does not distinguish between FPs and FNs. Therefore, the accuracy of a drop-out prediction algorithm does not indicate how often the model fails to identify an at-risk student. If drop-out is rare, the model could boast high accuracy without serving its intended purpose—such as helping administrators identify suitable students for intervention. Similarly, although summary metrics such as F_1 will appropriately penalize a model for its systematic failure to identify either positives or negatives, they do not provide transparent information to research consumers regarding the prevalence of these errors (Green & Viljoen, 2020). To improve the policy relevance of performance statistics, researchers can do the following:

- They can present metrics that transparently characterize the degree and types of errors (as in Arthur & Chang, 2024; Bird et al., 2021; Kelly et al., 2018). In binary classifiers, this includes precision, recall, specificity, and FP rate. In predicting graduation, for example, Bird et al. (2021) presented both precision (the share of true graduates that the model predicts will graduate) and recall (the share of predicted graduates who graduate). With a continuous classifier, metrics that clearly report the degree of error include raw error, mean squared error, and root mean squared error.
- They can present multiple performance metrics alongside each other. Bird et al. (2021), for example, provided bar charts to demonstrate that graduation recall was routinely higher than graduation precision.

Model Optimized to Testing Data

In calculating performance metrics, researchers commonly have two goals: selecting between competing algorithms and hyperparameters (model selection/tuning) and estimating the final model's performance (model evaluation). For final performance metrics to provide an unbiased estimate, however, these two functions must be completed on independent datasets. Otherwise, if testing data are used to support a choice between competing models, then final performance metrics will underestimate the true error, sometimes substantially (Hastie et al., 2009). Peeking repeatedly at testing statistics is akin to *p*-hacking; just as a quantitative researcher can exploit random statistical variation to inflate *p* values, a machine learning researcher can exploit random variation in the testing data to inflate performance metrics (Yarkoni & Westfall, 2017). To protect against this threat, researchers commonly can do the following:

- They can split labeled data into three datasets instead of two: training, development, and testing. This second split between training and development data can be used for model tuning and algorithm selection, whereas the testing data are only used once, after the model has been finalized (see, e.g., Lang et al., 2022).
- They can split labeled data into two overarching datasets, training and testing, but use *k*-fold validation within the training dataset to select the algorithm and hyperparameters (Hastie et al., 2009). In this approach, the training dataset is divided into *k* (commonly five or 10) equally sized subsets, or *folds*. The researcher trains the model *k* times, each time using *k* − 1 folds for training and the remaining fold for validation, rotating through all folds as the validation set. After identifying the best-performing hyperparameters and model setup, the model is retrained on the full training set before being validated on the hold-out testing dataset. See, for example, deployment of 10-fold validation by Bird et al. (2021).
- They can preregister the specifics of the machine learning training process (e.g., the cross-validation

method, hyperparameters tested, etc.), as demonstrated by Cimpian and Timmer (2019).

Dependence Between Training and Testing Data

For performance metrics to be unbiased, researcher decisions not only must be independent of the testing data, but the testing data itself also must be independent of the training data. In other words, knowing the outcome of an observation in the training dataset should not be useful for predicting the outcome of an observation in the testing dataset. This assumption is violated if, for example, the same person produced the two observations (e.g., one student produced an essay in the training dataset and another essay in the testing dataset) or if there are duplicates in the data (e.g., tweets that have been copy-pasted or retweeted by multiple users). If the degree of dependence is substantial, then a model could fit to noise in the dataset. Consider again the example of student essays. If one strong writer has an idiosyncratic writing style, the model might fit to those ungeneralizable idiosyncrasies. If that same writer has observations in the testing dataset, the model won't be penalized for such overfitting. To address this threat, researchers commonly can do the following:

- They can split the data for training and testing (or cross-validation) using the uppermost level of a hierarchical dataset. This might mean splitting observations at the person level (when individuals produce multiple observations), at the classroom level (when students are nested within classrooms), or at the school level (when teachers are nested within schools). For example, in training a classifier to identify authentic questions from teachers, Kelly et al. (2018) employed "leave one teacher out validation" so that performance metrics could not be overinflated via overfitting to individual teacher idiosyncrasies in the training data.

Insufficient Validation Sample Size

When researchers calculate performance metrics, these are point estimates derived from a sample (the testing data) with the purpose of generalizing to a population to which the model will be applied subsequently. As with all point estimates, these statistics should not be interpreted as the truth but rather as the best estimate of an unknown population parameter (Savoy, 1997). Thus, just as quantitative evaluation researchers present standard errors and confidence intervals regarding treatment-effect estimates, machine learning researchers should present confidence intervals surrounding performance metrics (Mitchell et al., 2019). Presenting confidence intervals would force researchers to acknowledge that high performance in the testing data, particularly in a small testing dataset, may be due to a lucky draw. Presenting confidence intervals also might encourage researchers to increase the size of their testing datasets, thereby increasing the statistical validity of their estimates. Approaches to confidence interval estimation include the following:

- Estimating a binomial proportion interval. In the case of binary predictions, researchers may estimate a confidence interval by calculating a binomial proportion interval: $\hat{p} \pm z\sqrt{\left[\hat{p}(1-\hat{p})\right]/n}$, where $\hat{p}$ is an estimated proportion-based performance metric (such as accuracy, recall, or precision), z is a critical value for a desired level of confidence, and n is the size of the data on which the metric is estimated. Consider this approach in the context of the work of Bird et al. (2021), for example. With an n of ~11,220 graduates (33,000 students in the testing sample × a graduation rate of 0.34), and a recall of 0.75, the estimated confidence interval surrounding recall for one of their prediction models would be approximately ± 0.008. If there were instead just 100 students in the testing sample (with an expected 34 graduates), the confidence interval surrounding recall would have been approximately ± 0.15.
- Bootstrapping the testing sample. For a given sample of n observations in a testing dataset $X = \{x_1, x_2, x_3, \ldots, x_n\}$, researchers would generate a set of bootstrap samples $X^{*i} = \{x_1^*, x_2^*, x_3^*, \ldots, x_k^*, \ldots, x_n^*\}$ for i through B using random sampling with replacement from X. Each bootstrap sample contains n members of the sample X, with some appearing zero times, some once, some twice, and so on. Within each bootstrapped sample, the researcher calculates the appropriate performance statistics (Savoy, 1997). The standard deviation of the resulting distribution is the bootstrapped standard error, and a 95% confidence interval can be obtained by assessing which two values 95% of the bootstrapped estimates fall between.

Internal Validity in Supervised Learning

Internally valid studies can help determine the extent to which an educational program has a positive impact on students, making it a top priority among governmental and funding agencies (What Works Clearinghouse, 2019). Contemporary education researchers thus are often highly attuned to methods that increase causal rigor. However, the growth of machine learning in education is somewhat at odds with a prioritization of internal validity. Although an algorithm will identify the combination of variables that best predicts the outcome of interest, there is no consideration of whether those variables are confounders of or contributors to the outcome. Further, there is no guarantee that the individual variables that are given the greatest weight in the model are the same variables that are most predictive of the outcome—only that the combination of variables is maximally predictive (Mullainathan & Spiess, 2017). Quite simply, supervised learning algorithms are optimized for prediction rather than causal inference. Although experiments and quasi-experiments are designed to estimate the impact of *A* on *B*, supervised learning methods are designed to estimate *predictions* of *B* from *A* (Mullainathan & Spiess, 2017).

Nevertheless, prediction can be used in the service of causal inference. Three common scenarios were identified from the reviewed literature. First, supervised learning algorithms may be used to measure outcomes or characterize treatments in an evaluation framework (as in Anglin, 2024; Harper et al., 2021; Mozer et al., 2023). Second, supervised learning algorithms may be used to build causal theory, particularly surrounding moderators, or to estimate heterogeneous treatment effects (as in Master et al., 2022; Pietsch et al., 2023; Suk & Han, 2024). Third, supervised learning algorithms are increasingly being used to identify and control for confounds (as in Gormley et al., 2023; Jabbari et al., 2023; Keller, 2020).

In the first case, where supervised learning is used to measure treatments or outcomes, the same threats to internal validity that might occur with any evaluation apply, including ambiguous temporal precedence, selection, history, maturation, regression, attrition, testing, instrumentation, and the additive and interactive effects of these (Shadish et al., 2002). Although a comprehensive overview of these threats is beyond the scope of this paper, readers may look to the Registry of Educational Effectiveness (Spybrook et al., 2019) and the What Works Clearinghouse (2019) protocols. This section highlights threats relevant to the second and third cases.

Instability and Selection Bias in Predictor Importance

Machine learning algorithms are adept at identifying nonlinear and interactive patterns in data (Hastie et al., 2009). They are thus especially useful for identifying heterogeneity in phenomena; researchers may use supervised learning to predict an outcome (e.g., graduation) and then observe the variables that are most predictive—such as the largest coefficients in a penalized regression or the first branches in a regression tree—to build causal theory around the variables that increase or decrease the outcome. If there are important interactive and nonlinear relations—for example, if men in STEM majors are at the greatest risk of dropping out or if a precipitous, rather than linear, drop in GPA causes students to leave school—supervised learning models can efficiently identify these patterns, helping researchers to build inductive theory (Choudhury et al., 2018). However, there are challenges in this approach. First, the most important predictor in a given model is not necessarily the most important available predictor of the outcome. Due to the flexibility of many supervised learning algorithms, slight variations in training data can cause notable changes in predictor importance, even while model performance remains unchanged (Keller, 2020; Mullainathan & Spiess, 2017). For this reason, the variables identified as highly predictive using flexible and adaptive algorithms such as regression trees and gradient boosting are less stable than those identified using ordinary least squares regression (Mullainathan & Spiess, 2017).

Furthermore, as with any analysis of patterns in observational data, a variable may be a stable and significant predictor of an outcome without necessarily having a causal impact on it. The identified predictor simply may be a correlate of another, unobserved variable—the true determinant. For example, a hypothetical supervised learning model may find that undergraduate students in a particular major are more likely to graduate. This may be due to their experiences in the major (i.e., a causal relationship) or because of the type of student who decides to pursue the major (i.e., selection bias). The model will not distinguish between these two possibilities. To address these threats when building theory, researchers may do the following:

- Acknowledge that findings regarding predictor importance are correlational and exploratory (as in Lang et al., 2022; Master et al., 2022).
- Use supervised learning to identify potentially important predictors and then assess the predictor-outcome relationship in a separate hold-out dataset, addressing the challenge of predictor instability. This is the approach taken by Master et al. (2022) when identifying potential moderators of principal coaching effects—training a causal forest on one portion of the data and then using a hold-out dataset to assess moderator importance.
- Assess average predictor importance across many models (as in González Canché, 2023; Master et al., 2022). In an ensemble approach to supervised learning, a researcher trains many models on random subsets of the data—combining many regression trees into a forest, for example. Final predictions then result from aggregation across the models. Just as predictions are more stable in ensemble models, predictor importance is also more stable when aggregating across several models (Elith et al., 2008).

Unobserved Confounders in Models Predicting Treatment Selection

Finally, a common application of supervised learning in causal research is to aid the identification and control of confounders. For example, researchers commonly use regression trees to predict treatment take-up (McCaffrey et al., 2004). The resulting predicted probability scores are then used in a propensity score framework to control for selection. Empirical evidence from the within-study comparison literature suggests that—given the same set of potential covariates—machine learning approaches to propensity score estimation can reduce bias when compared with logistic regression approaches (Anglin et al., 2023). However, as with any matching or weighting approach, the algorithm's success at eliminating selection bias depends on the quality of available data (Cook et al., 2008). Supervised learning cannot address the problem of unobserved confounders. To address the threat of unobserved confounders, researchers commonly can do the following:

- They can present evidence of similarity between the treatment and comparison groups following propensity score weighting (as in Gormley et al., 2023; Im et al., 2016; Sales et al., 2018). Although balance on observable characteristics does not guarantee balance on unobservable characteristics, a discernible imbalance does increase selection bias concerns.
- They can collect data on hypothesized predictors of treatment take-up. Selection bias is often reduced substantially when researchers control for pretreatment outcome measures and for variables that are theorized to influence selection, such as motivation or preferences (Keller, 2020; Marcus et al., 2012; Pohl et al., 2009; Wong et al., 2017). In contrast, exclusively controlling for demographic covariates rarely produces unbiased treatment effects (Wong et al., 2017).

Research Protocol

Drawing on the threats and best practices described earlier, the research protocol presented in Table 4 provides an initial starting point for improving and assessing the validity of inferences drawn from machine learning applications. Like the validity-types framework, the protocol emphasizes *proactive* design decisions. By considering threats during the planning stages of a study, researchers may preemptively address them—a sentiment often captured by the adage, "You can't fix with analysis what you've bungled by design" (Light et al., 1990, p. xiii). Researchers can best address construct validity by identifying the construct of interest upfront and by selecting training data that best reflect that construct. They can best address external validity by ensuring that the training and testing sample and setting match the context(s) where the model likely will be applied and by ensuring the adequate representation of population subgroups. They can best address statistical validity by selecting the most informative performance metrics and by ensuring an adequate sample size in the testing data. And they can best address internal validity by selecting an appropriate design and by collecting data on the most relevant confounders. The protocol provided in Table 4 prompts researchers to consider these facets in the early stages of a study.

The validity of supervised learning applications also may be increased post hoc (i.e., after model training) through comprehensive reporting and transparency (Gebru et al., 2021; Mitchell et al., 2019). In the machine learning literature, the push for increased transparency has involved the increased adoption of standardized documentation to accompany public-use training datasets (Gebru et al., 2021) and pretrained models (Mitchell et al., 2019). Although studies applying supervised learning in educational contexts rarely release their training data or models, this approach is nonetheless instructive. To judge the validity of inferences drawn from supervised learning, critical readers require comprehensive information. To this end, the questions in Table 4 may serve as a prompt for future study authors when deciding which information to include in a paper.

Discussion and Limitations

This article draws a parallel between validity typology of Shadish et al. (2002) and the inferences drawn from supervised learning in educational contexts. It provides a holistic overview of threats to validity alongside example approaches for addressing those threats. The article's aim is to improve the validity of supervised learning applications in education research. Naturally, however, its limitations reflect both the limitations of the original typology and of typologic approaches more generally.

First, catalogues of validity types and threats serve as heuristics for researchers (Mark, 1986). That is, these threats represent cognitive shortcuts (Reichardt, 1985). A catalog of various threats allows us to evaluate the validity of inferences more easily than we otherwise might (Mark, 1986), particularly given the heavy cognitive lift required to evaluate the validity of inferences derived using unfamiliar methods. However, such shortcuts also may serve as blinders, allowing unlisted threats to go unacknowledged (Reichardt, 1985). Further, typologies suffer from an inherent arbitrariness. Critics have pointed out that "definitions of validity and threats to validity have varied over time, are sometimes incongruous, and are not always easy to differentiate" (Reichardt, 2019, p. 27). As Mark (1986, p. 63) writes, "A validity typology is not a foolproof, logistically consistent, mutually exclusive set of categories. It is a device, an aid." Even if distinctions between validity types and threats remain up for debate, therefore, attempts to collate and organize them still can prove valuable.

Second, any list of threats necessarily will be incomplete. Indeed, the number of threats identified by Shadish et al. (2002) tripled between 1957 and 1979 (Campbell, 1957; Cook & Campbell, 1979). The threats identified in this article are thus not expected to be comprehensive. Although machine learning applications in education are increasing quickly, the literature base is still relatively young; new challenges likely will be identified as the field

TABLE 4

Summary Protocol for Machine Learning Applications in Education

Questions	CV	EV	SV	IV
What are the key research questions and hypotheses?	X	X	X	X
What role do machine learning models play in the study?	X	X	X	X
Define the construct(s) you aim to measure with a machine learning model, and link the conceptualization to prior literature.	X			
To what extent is there slippage between the construct of interest and the labels in the data? If the labels assigned to the data differ from the construct of interest, describe how.	X			
If gold-standard data involve labels assigned by human coders, what were the specific instructions and materials provided to the labeler(s)?	X			
Describe the labelers' training and experience.	X			
How will you measure inter-rater agreement?	X			
Describe the predictors you will allow your model to consider.	X			
Which of these are likely correlates of a confounding construct?	X			
How, if at all, will you observe predictor importance?	X			
Will there be more than one measure of the construct of interest? If so, is at least one of these measures not reliant on machine learning?	X			
Do your participants have the means and/or motivation to game the model?	X			
If so, how do you plan to probe participants' reactions to the model?	X			
If participants were to game the model, would this behavior be positive, negative, or neutral for student learning?	X			
What is the target population for your model?		X		
Describe the source of your training and testing datasets. To what extent is there theoretical alignment and misalignment with the target population and the sample population?		X		
Describe your proposed sample with respect to subgroups (e.g., what proportion of your population has an individualized educational plan)?		X		
For which subgroups will you report performance statistics?		X		
If you will be using a pretrained model, how will you validate the model in its current setting?		X		
Over what time period will your model be employed? Is the full period represented in your training and testing data?		X		
To what extent do you expect the predictive capability of the model's features to change during the model's employment period?		X		
Can you empirically assess model drift by assessing changes in performance over time?		X		
What are the most relevant performance metrics?			X	
What is the size of your labeled dataset?			X	
What is the intended training/development/testing ratio?			X	
What is the count of true positives and true negatives in the testing data?			X	
Records may be unintentionally recorded twice. How will you assess your data for possible duplicates?			X	
If the data are nested, describe the nesting structure and the level at which you will split your data for training/testing?			X	
How will you protect against the temptation to peek at your testing data?			X	
How will you report uncertainty around your performance metrics?			X	
What, if any, causal inferences are embedded within the research question?				X
What design features are included in the study to address threats to internal validity (e.g., selection bias, time-varying confounders)? See the Registry for Educational Effectiveness studies for in-depth guiding questions relevant to your chosen research design (Anderson et al., 2019).				X

Note. An X indicates the most relevant validity type to which the question speaks. CV = construct validity; EV = external validity; SV = statistical validity; IV = internal validity.

develops. Similarly, best practices are also likely to grow and evolve, meaning that the approaches discussed here and in the protocol should be considered as examples rather than as a comprehensive list of requirements.

Third, Shadish et al. (2002) may themselves take issue with the application of their validity typology to supervised learning applications. These authors have long argued that internal validity is the sine qua non of research; in their view, internal validity must be prioritized before assessments regarding other validity types are deemed appropriate (Campbell & Stanley, 1963). In contrast, overemphasizing internal validity at the cost of other validity types has been heavily critiqued in discussions of the original validity typology (Albright & Malloy, 2000; Reichardt, 2019). This article is thus not the first to advocate for expanding the validity typology to include noncausal research (Huck & Sandler, 1979; McMillan, 2000; Onwuegbuzie, 2000).

Finally, as noted earlier, the understanding of validity provided by Shadish et al. (2002) is only one formulation among many and is not without its limitations. One key drawback of the framework, when applied to supervised learning, is the relatively limited focus it places on *consequences* and *value implications* (Kane, 2001; Messick, 1989). The threats to validity given by Shadish et al. (2002) focus on the *causes* of faulty inferences, encouraging researchers to rule out these threats and improve their inferences. However, comparatively less attention is given to the consequences of these inferences. As Kane (2001) points out, even accurate inferences are not sufficient to argue for test use; a medical test that can accurately predict an untreatable disease may still cause harm if applied without purpose, particularly if there are side effects. Similarly, even an accurate supervised learning model may have unintended consequences when applied in practice (Barocas et al., 2023; see also Lee et al. [2021] for an example of negative consequences resulting from a machine learning measure in higher education). Further, Shadish et al. (2002) only provide limited discussions of trust and transparency issues, key issues in supervised learning given that training datasets are rarely described and commonly underrepresent key demographic groups (Buolamwini & Gebru, 2018). For these reasons, the validity typology and associated checklists presented here cannot serve as the final conceptualization of machine learning validity in education research. Instead, they offer a practical form of scaffolding while best practice in the field develops.

Conclusions

Given the exponential rise of machine learning applications in education research, we are at a critical disciplinary juncture. Machine learning is equally capable of generating valuable insights and faulty inferences. This article aimed to increase the likelihood of the former by providing education researchers with a straightforward reference guide to validity considerations. Although machine learning technologies are quick to adapt and evolve, the most important questions concerning valid inferences are age old: Does the measured construct align with the construct's theoretical definition? Does the sample genuinely reflect the populations of interest? Are the statistics unbiased? Do the correlations reflect causation? This article encourages researchers to pay close attention to these facets of supervised learning applications, increasing their rigor even as they employ cutting-edge algorithms.

Declaration of Conflicting Interests

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kylie Anglin https://orcid.org/0000-0001-7661-3370

References

AERA. (2024). *Who we are*. https://www.aera.net/About-AERA/Who-We-Are

Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, Campbell, Cronbach, and enduring issues. *Review of General Psychology*, *4*(4), 337–353. https://doi.org/10.1037/1089-2680.4.4.337

Anderson, D., Spybrook, J., & Maynard, R. (2019). REES: A registry of efficacy and effectiveness studies in education. *Educational Researcher*, *48*(1), 45–50. https://doi.org/10.3102/0013189x18810513

Anglin, K. L. (2024). The role of state education regulation: Evidence from the Texas Districts of Innovation statute. *Educational Evaluation and Policy Analysis*, *46*(2), 534–554. https://doi.org/10.3102/01623737231176509

Anglin, K. L., Boguslav, A., & Hall, T. (2022). Improving the science of annotation for natural language processing: The use of the single-case study for piloting annotation projects. *Journal of Data Science*, *20*(3), 339–357. https://doi.org/10.6339/22-JDS1054

Anglin, K. L., Wong, V. C., Wing, C., Miller-Bains, K., & McConeghy, K. (2023). The validity of causal claims with repeated measures designs: A within-study comparison evaluation of differences-in-differences and the comparative interrupted time series. *Evaluation Review*, *47*(5), 895–931. https://doi.org/10.1177/0193841X231167672

Arthur, D., & Chang, H.-H. (2024). DINA-BAG: A bagging algorithm for DINA model parameter estimation in small samples. *Journal of Educational and Behavioral Statistics*, *49*(3), 1–16. https://doi.org/10.3102/10769986231188442

Aulck, L., Malters, J., Lee, C., Mancinelli, G., Sun, M., & West, J. (2021). Helping students FIG-ure it out: A large-scale study of freshmen interest groups and student success. *AERA Open*, *7*, 1–19. https://doi.org/10.1177/23328584211021857

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bergstrom, C. T., & West, J. D. (2021). *Calling bullsh*t: The art of skepticism in a data-driven world*. Random House Trade Paperbacks.

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open*, *7*, 1–19. https://doi.org/10.1177/23328584211037630

Bloom, H., & Michalopoulos, C. (2013). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups. *Prevention Science*, *14*(2), 179–188. https://doi.org/10.1007/s11121-010-0198-x

Bowyer, K. W., King, M. C., Scheirer, W. J., & Vangara, K. (2020). The "criminality from face" illusion. *IEEE Transactions on Technology and Society*, *1*(4), 175–183. https://doi.org/10.1109/tts.2020.3032321

Bradley, C. L., & Renzulli, L. A. (2011). The complexity of non-completion: Being pushed or pulled to drop out of high school. *Social Forces*, *90*(2), 521–545. https://doi.org/10.1093/sf/sor003

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Brown, T. M., & Rodríguez, L. F. (2009). School and the co-construction of dropout. *International Journal of Qualitative Studies in Education*, *22*(2), 221–242. https://doi.org/10.1080/09518390802005570

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. *PLMR*, *81*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental design for research* (p. 84). Houghton Mifflin.

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal*, *59*(6), 1122–1156. https://doi.org/10.3102/00028312221106773

Chi, O. L., & Lenard, M. A. (2023). Can a commercial screening tool help select better teachers? *Educational Evaluation and Policy Analysis*, *45*(3), 530–539. https://doi.org/10.3102/01623737221131547

Choudhury, P., Allen, R., & Endres, M. (2018). Developing theory using machine learning methods. *SSRN*. https://ssrn.com/abstract=3251077 or https://doi.org/10.2139/ssrn.3251077

Cimpian, J. R., & Timmer, J. D. (2019). Large-scale estimates of LGBQ–heterosexual disparities in the presence of potentially mischievous responders: A preregistered replication and comparison of methods. *AERA Open*, *5*(4), 1–35. https://doi.org/10.1177/2332858419888892

Cohen, J., Hutt, E., Berlin, R., & Wiseman, E. (2022). The change we cannot see: Instructional quality and classroom observation in the era of Common Core. *Educational Policy*, *36*(6), 1261–1287. https://doi.org/10.1177/0895904820951114

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Rand McNally.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750. https://doi.org/10.1002/pam

Cope, B., & Kalantzis, M. (2016). Big data comes to school: Implications for learning, assessment, and research. *AERA Open*, *2*(2), 1–19. https://doi.org/10.1177/2332858416641907

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281. https://doi.org/10.1037/h0040957

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 483–505. https://doi.org/10.3102/01623737231169270

Doran, H. (2023). A collection of numerical recipes useful for building scalable psychometric applications. *Journal of Educational and Behavioral Statistics*, *48*(1), 37–69. https://doi.org/10.3102/10769986221116905

Doroudi, S. (2020). The bias–variance tradeoff: How data science can inform educational debates. *AERA Open*, *6*(4), 1–18. https://doi.org/10.1177/2332858420977208

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Emmert-Streib, F., & Dehmer, M. (2019). Evaluation of regression models: Model assessment, model selection and generalization error. *Machine Learning and Knowledge Extraction*, *1*(1), 521–551. https://doi.org/10.3390/make1010032

Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. In: A. L. C. Bazzan & S. Labidi (Eds.), *Advances in Artificial Intelligence—SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence, Sao Luis, Maranhao, Brazil, September 29-Ocotber 1, 2004. Proceedings* (*Vol. 17*, pp. 286–295). Springer Nature Link. https://doi.org/10.1007/b100195

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). *Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?* ACM Digital Library. https://doi.org/10.1145/3351095.3372862

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

González Canché, M. S. (2023). The geography of mathematical (dis) advantage: An application of multilevel simultaneous autoregressive (MSAR) models to public data in education research. *AERA Open*, *9*(1), 1–38. https://doi.org/10.1177/23328584231198

Gormley, W. T., Jr., Amadon, S., Magnuson, K., Claessens, A., & Hummel-Price, D. (2023). Universal pre-K and college enrollment: Is there a link? *AERA Open*, *9*(1), 1–17. https://doi.org/10.1177/23328584221147893

Green, B., & Viljoen, S. (2020, January 27–30). *Algorithmic realism: Expanding the boundaries of algorithmic thought*. ACM Digital Library. https://doi.org/10.1145/3351095.3372840

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 1–31. https://doi.org/10.1093/pan/mps028

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361. https://doi.org/10.3102/1076998619832248

Harper, D., Bowles, A. R., Amer, L., Pandža, N. B., & Linck, J. A. (2021). Improving outcomes for English learners through technology: A randomized controlled trial. *AERA Open*, *7*(1), 1–20. https://doi.org/10.1177/23328584211025528

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. In P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, & S. Zeger (Eds.), *Spring series in statistics* (2nd ed., p. 694). Springer. http://www.springerlink.com/index/10.1007/b94608

Hovy, D., & Spruit, S. L. (2016). *The social impact of natural language processing*. ACL Anthology. https://doi.org/10.18653/v1/P16-2096

Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses: Alternative interpretations of data based conclusions*. Harpercollins College Division.

Im, M. H., Hughes, J. N., Cao, Q., & Kwok, O. (2016). Effects of extracurricular participation during middle school on academic motivation and achievement at grade 9. *American Educational Research Journal*, *53*(5), 1343–1375. https://doi.org/10.3102/0002831216667479

Jabbari, J., Chun, Y., Huang, W., & Roll, S. (2023). Disaggregating the effects of STEM education and apprenticeships on economic mobility: Evidence from the LaunchCode program. *Educational Evaluation and Policy Analysis*, 1–24. https://doi.org/10.3102/01623737231199985

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527. https://doi.org/10.1037//0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine learning-based science. *Patterns*, *4*(9), 1–24. https://doi.org/10.1016/j.patter.2023.100804

Keller, B. (2020). Variable selection for causal effect estimation: Nonparametric conditional independence testing with random forests. *Journal of Educational and Behavioral Statistics*, *45*(2), 119–142. https://doi.org/httpsL//doi.org/10.3102/1076998619872001

Kelly, S., & Abruzzo, E. (2021). Using lesson-specific teacher reports of student engagement to investigate innovations in curriculum and instruction. *Educational Researcher*, *50*(5), 306–314. https://doi.org/10.3102/0013189X2098225

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464. https://doi.org/10.3102/0013189X18785613

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085x.2021.1927212

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Lang, D., Wang, A., Dalal, N., Paepcke, A., & Stevens, M. L. (2022). Forecasting undergraduate majors: A natural language approach. *AERA Open*, *8*(1), 1–18. https://doi.org/10.1177/23328584221126516

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lee, J. C., Dell, M., González Canché, M. S., Monday, A., & Klafehn, A. (2021). The hidden costs of corroboration: Estimating the effects of financial aid verification on college enrollment. *Educational Evaluation and Policy Analysis*, *43*(2), 233–252. https://doi.org/10.3102/0162373721989304

Li, X., Xu, H., Zhang, J., & Chang, H. (2023). Deep reinforcement learning for adaptive learning systems. *Journal of Educational and Behavioral Statistics*, *48*(2), 220–243. https://doi.org/10.3102/10769986221129847

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Harvard University Press.

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614. https://doi.org/10.3102/01623737211009267

Lucy, L., Demszky, D., Bromley, P., & Jurafsky, D. (2020). Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks. *AERA Open*, *6*(3), 1–27. https://doi.org/10.1177/2332858420940312

Lwakatare, L. E., Raj, A., Crnkovic, I., Bosch, J., & Olsson, H. H. (2020). Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. *Information and Software Technology*, *127*, 106368. https://doi.org/10.1016/j.infsof.2020.106368

Manning, C. D., Schütze, & Hinrich. (1999). *Foundations of statistical natural language processing* (p. 680). MIT Press. https://dl.acm.org/citation.cfm?id=311445

Marcus, S. M., Stuart, E. a., Wang, P., Shadish, W. R., & Steiner, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods*, *17*(2), 244–254. https://doi.org/10.1037/a0028031

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentation. *New Directions for Program Evaluation*, *1986*(31), 47–66. https://doi.org/10.1002/ev.1433

Master, B. K., Schwartz, H., Unlu, F., Schweig, J., Mariano, L. T., Coe, J., Wang, E. L., Phillips, B., & Berglund, T. (2022). Developing school leaders: Findings from a randomized control trial study of the Executive Development Program and paired coaching. *Educational Evaluation and Policy Analysis*, *44*(2), 257–282. https://doi.org//10.3102/01623737211047256

McCaffrey, D. F., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Mcfarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. *AERA Open*, *7*(1), 1–12. https://doi.org/10.1177/23328584211052055

McMillan, J. H. (2000). *Examining categories of rival hypotheses for educational research*. Annual Meeting of the American Educational Research Association, New Orleans. https://files.eric.ed.gov/fulltext/ED447194.pdf

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189x018002005

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting*. ACM Digital Library. https://doi.org/10.1145/3287560.3287596

Mozer, R., Miratrix, L., Relyea, J. E., & Kim, J. S. (2023). Combining human and automated scoring methods in experimental assessments of writing: A case study tutorial. *Journal of Educational and Behavioral Statistics*, *49*(5), 780–816. https://doi.org/10.3102/10769986231207886

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

National Center for Special Education Research (NCSER). (2021, November). *AI-augmented learning for individuals with disabilities: New funding opportunity, current research, and the potential for improving student outcomes*. Inside IES Research. https://ies.ed.gov/blogs/research/post/ai-augmented-learning-for-individuals-with-disabilities-new-funding-opportunity-current-research-and-the-potential-for-improving-student-outcomes

Nystrand, M. (2004). *CLASS 4.0 user's manual*. National Research Center on English Learning and Achievement. https://class.wceruw.org/documents/class/CLASS%204%20Documentation.pdf

Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue*. Teachers College Press.

Onwuegbuzie, A. J. (2000). *Expanding the framework of internal and external validity in quantitative research*. Annual Meeting of the Association for the Advancement of Educational Research, Ponte Verde, FL. https://files.eric.ed.gov/fulltext/ED448205.pdf

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *Aera Open*, *3*(4). https://doi.org/10.1177/2332858417749220

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pang, B., Nijkamp, E., & Wu, Y. N. (2020). Deep learning with tensorflow: A review. *Journal of Educational and Behavioral Statistics*, *45*(2), 227–248. https://doi.org/10.3102/1076998619872761

Phelps, R. P. (2011). *Standards for educational & psychological testing*. American Psychological Association.

Pietsch, M., Aydin, B., & Gümüş, S. (2023). Putting the instructional leadership–student achievement relation in context: A meta-analytical big data study across cultures and time. *Educational Evaluation and Policy Analysis*, 1–36. https://doi.org/10.3102/01623737231197434

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*(4), 463–479. https://doi.org/10.3102/0162373709343964

Ramirez, G., Hooper, S. Y., Kersting, N. B., Ferguson, R., & Yeager, D. (2018). Teacher math anxiety relates to adolescent students' math achievement. *AERA Open*, *4*(1), 1–13. https://doi.org/doi.org/10.1177/2332858418756052

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In D. Grusky (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (*Vol. 1*, pp. 91–116), Routledge.

Reardon, S. F., & Stuart, E. A. (2019). Education research in a new data environment: Special issue introduction. *Journal of Research on Educational Effectiveness*, *12*(4), 567–569. https://doi.org/10.1080/19345747.2019.1685339

Reichardt, C. S. (1985). Reinterpreting Seaver's (1973) study of teacher expectancies as a regression artifact. *Journal of Educational Psychology*, *77*(2), 231–236. https://doi.org/10.1037/0022-0663.77.2.231

Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford Press.

Rosenberg, J. M., Borchers, C., Dyer, E. B., Anderson, D., & Fischer, C. (2021). Understanding public sentiment about educational reforms: The Next Generation science standards on Twitter. *AERA Open*, *7*(1), 1–17. https://doi.org/10.1177/23328584211024261

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in behavioral research* (pp. 280–350). Oxford University Press.

Rothacher, Y., & Strobl, C. (2024). Identifying informative predictor variables with random forests. *Journal of Educational and Behavioral Statistics*, *49*(4), 595–629. https://doi.org/10.3102/10769986231193327

Sales, A. C., Hansen, B. B., & Rowan, B. (2018). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 3–31. https://doi.org/10.3102/1076998617731518

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, *33*(4), 495–512. https://doi.org/10.1016/S0306-4573(97)00027-7

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., & Young, M. (2014). *Machine learning: The high-interest credit card of technical debt*. Google Research.

https://research.google/pubs/machine-learning-the-high-interest-credit-card-of-technical-debt/

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shaffer, D. W., & Ruis, A. R. (2021). How we code. In S. B. Lee & A. R. Ruis (Eds.), *ICQE 2020*. Springer.

Shores, K., & Steinberg, M. P. (2022). Fiscal federalism and K–12 education funding: Policy lessons from two educational crises. *Educational Researcher*, *51*(8), 551–558. https://doi.org/10.3102/0013189X221125764

Si, Y., Little, R. J., Mo, Y., & Sedransk, N. (2023). A case study of nonresponse bias analysis in educational assessment surveys. *Journal of Educational and Behavioral Statistics*, *48*(3), 271–295. https://doi.org/10.3102/10769986221141074

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106. https://doi.org/10.3102/1076998616666808

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). *Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks*. ACL Anthology. https://aclanthology.org/D08-1027

Spybrook, J., Anderson, D., & Maynard, R. (2019). The Registry of Efficacy and Effectiveness Studies (REES): A step toward increased transparency in education. *Journal of Research on Educational Effectiveness*, *12*(1), 5–9. https://doi.org/10.1080/19345747.2018.1529212

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135–153. https://doi.org/10.3102/1076998609359791

Suk, Y., & Han, K. T. (2024). A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*, *49*(2), 151–172. https://doi.org/10.3102/10769986231171711

Suk, Y., Kim, J.-S., & Kang, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *Journal of Educational and Behavioral Statistics*, *46*(3), 323–347. https://doi.org/10.3102/1076998620951983

Suresh, H., & Guttag, J. (2021). *A framework for understanding sources of harm throughout the machine learning life cycle*. Equity and Access in Algorithms, Mechanisms, and Optimization. https://doi.org/10.1145/3465416.3483305

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*, 319–330. https://doi.org/10.28945/331

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789

What Works Clearinghouse. (2019). *What works clearinghouse standards handbook: Version 4. U.S* (pp. 1–17). Department of Education's Institute of Education Sciences (IES). https://doi.org/10.1037/e578392011-004

Wong, V. C., Valentine, J., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness*, *10*(1), 207–236. https://doi.org/10.1080/19345747.2016.1164781

Wu, E., & Gagnon-Bartsch, J. A. (2021). Design-based covariate adjustments in paired experiments. *Journal of Educational and Behavioral Statistics*, *46*(1), 109–132. https://doi.org/10.3102/1076998620941469

Wu, X., & Zhang, X. (2016). *Automated inference on criminality using face images*. https://www.researchgate.net/publication/310235081_Automated_Inference_on_Criminality_using_Face_Images

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zadrozny, B. (2004). *Learning and evaluating classifiers under sample selection bias*. ACM Digital Library. https://doi.org/10.1145/1015330.1015425

Zheng, Y., Nydick, S., Huang, S., & Zhang, S. (2024). MxML (exploring the relationship between measurement and machine learning): Current state of the field. *Educational Measurement: Issues and Practice*, *43*(1), 19–38. https://doi.org/10.1111/emip.12593

Author

KYLIE ANGLIN is an assistant professor in research methods, measurement, and evaluation at the University of Connecticut. Her research leverages machine learning and natural language processing to advance research methodologies.