13 A tale of two worlds: Machine learning approaches at the intersection with educational measurement

By Kathleen Scalise, Cassandra Malcom and Errol Kaylor

(University of Oregon)

Promising digital technology affordances have expanded rapidly in education, and advances in the volume and nature of evidence that can be generated through digital technologies are impressive. However, especially at scale, analytical approaches to accumulate such data and draw meaningful conclusions (inferences) remain a frontier that is hard to navigate. This chapter discusses how machine learning and artificial intelligence approaches rapidly emerging in educational contexts are intersecting in many ways with educational measurement and argues for the imperative of these different fields to learn from each other. The chapter suggests some main takeaways for each field for the valid use and interpretation of innovative educational assessments.

Introduction

This publication makes the case that large-scale assessments should move beyond what is easy to assess and develop new methods to make claims on students' development of complex competencies. It also argues that simulations and other interactive assessment experiences can supplement more traditional item formats and potentially provide the evidence we need to compare students on these competencies. Scholars in learning analytics (LA) have made tremendous progress in applying data mining (DM) and machine learning (ML) techniques to the streams of data generated from learning experiences in digital environments. It is thus tempting to imagine that all we need to innovate assessment is to adopt these new techniques, moving beyond established methods of psychometrics. However, things are not so straightforward because the bridge between learning analytics and educational measurement still needs to be built.

The gap that exists between the two fields can be explained by their principal goals. The goal of LA is often to describe how learners learn or to find ways to adapt and personalise learning content to individual learners – sometimes also called Learning Engineering. LA uses a variety of data-driven approaches to make predictions about how people will learn in specific situations (i.e. using certain tools or working on certain activities). On the other hand, educational measurement – sometimes called psychometrics – focuses on making defensible claims about students' achievement, abilities or engagement with learning. Both can be used to inform learning interventions. LA and educational measurement are not completely distinct worlds: if the evidence collected for LA is then used to generate metrics (i.e. to make claims about students or student groups), then educational measurement is involved *even if the end result is framed as a prediction rather than a score report*. Some researchers including Sclater (2014[1]) and Wilson and Scalise (2016[2]) have therefore begun to establish standards of practice in LA when educational measurement is involved.

Hidden challenges for accumulating complex evidence

LA developers as well as those working in educational measurement have been converging in recent years, at least in very basic ways, on some important topics at the interface of LA and measurement technology. However, distance remains because scholars across these fields have different scholarly preparation, discourse language, epistemologies, ontological commitments and pedagogical grammars. For instance, in LA, conclusions tend to hinge on a relative argument about which model better explains the data set and is therefore the better "predictor" of something that the data set is purported to represent. This makes a lot of sense for traditional machine learning fields that rely on manifest variables (i.e. something that can be directly measured or observed) and involve data mining of homogeneous data sets – for example, a computer vision application that identifies if there is a tiger in a given picture (the tiger being the manifested variable). Yet for latent traits, such as mathematics knowledge and mental health, manifest variables do not exist and therefore there remain questions about what the data set represents.

Given these differences, LA scholars often see the need for new content and technology affordances in educational measurement but are unfamiliar with addressing what we refer to as the "Levy challenges". Over several decades, the field of psychometrics has developed well-accepted procedures for important issues in educational measurement, which include calibration and estimation of overall claims, reliability and precision information, test form creation, linking and equating, adaptive administrations, evaluating assumptions, checking data-model fit, differential functioning and invariance. However, as argued by Levy (2012[3]) and others, newer "data analytic" techniques using machine learning cannot rely on the same well-fitting measurement models used in psychometrics to verify the quality of assessment evidence and establish validity – an issue we call the "Levy challenges" and introduce in Chapter 8 of this report. LA approaches currently lack the theory and methods as well as the operational infrastructure often needed, such as analytic programs and delivery platforms at scale, to address these challenges. What this means

is that the way we can and should interpret findings about learners changes considerably if methods of evidence accumulation cannot satisfy these challenges via well-accepted procedures.

It should also be noted that classifications by software, such as describing how learners learn by classifying them into categories, also involves indicators and inferences. There would be no way to classify the learners if there were no indicators, and no reason to classify if no inferences were to be made from these classifications. Inferences might be about what might help the learner learn, but regardless, such inferences are claims about the student.

Many therefore advocate for a "separate-but-equal" view to break out different types of assessments by purpose. This sounds promising, but here is the resulting dilemma: if assessments for different purposes are treated as "separate-but-equal," this makes it simpler to analyse and implies no need to make consistent claims based on learning analytics methods and claims based on educational measurement – but this compromises the utility of assessments. Inconsistent claims across assessments – even if those assessments have different purposes (e.g. to plan future learning vs. to summarise what has been learned) – are highly confusing to teachers, students, parents and policy makers. For example, if teachers are told with one analytic technique that their students are proficient in their science knowledge while another technique says the same students have large learning gaps, teachers will not know what to believe and may ultimately discount the use of evidence in their work.

One solution that some suggest is to advance the field of assessment design by administering complex technology-based tasks at scale to generate process data but refrain from using the data for making inferential claims in the reporting (instead leaving these rich data for secondary research). However, such tasks may then be quickly discounted for various reasons: for example, is it a legitimate use of resources to develop technology-rich tasks if the new evidence they generate is not then used for reporting? Is it a legitimate use of student, school and administrators time to sit them? If such tasks are needed to fully actualise the construct, how can leaving out the evidence be justified? If the data are to be used for research, should the research not happen first and then deployment at scale occur when evidence is ready for reporting?

Another possibility is to neglect the "Levy challenges" altogether and be content with new analytic and reporting approaches that haven't yet matured robust measurement procedures. But as discussed throughout this report and elsewhere, the validity issues associated with integrating complex data into measures can be substantial. Without the ability to calibrate and estimate overall score(s), generate reliability and precision information, conduct subgroup analyses, create test forms, and engage in linking and equating, are robust inferences from large-scale assessments possible? Adaptive administrations, evaluating assumptions across languages and cultures, checking data-model fit, investigating differential functioning, and establishing invariance in the context of hard-to-measure constructs and complex naturalistic tasks goes well beyond what is currently known (Scalise, 2012[4]).

The intersection of the two worlds

As explored in more detail in Chapter 8 of this report, possible approaches to accumulating complex pieces of evidence to make inferences about students in large-scale assessment hold some promise. These include developing extended measurement models that borrow strengths from both psychometrics and learning analytics, establishing multiple inferential grain sizes or iteratively developing exploratory models to ultimately reach a confirmatory one. When considering any approach to accumulating (and eventually, reporting) assessment data, it is important to be mindful of the intended purpose of the assessment and of how results will be used and what they might be interpreted to mean. Given the hidden challenges described above, in the rest of this chapter we discuss some things that the fields of LA and educational measurement might learn from each other to further innovation in assessment analytics – especially in the context of large-scale assessments.

What learning analytics can learn from educational measurement

Evidence accumulation is as important as evidence elicitation

We argue that perhaps the most important thing that LA needs to learn about educational measurement is that when techniques such as machine learning (ML) – and more complex types of machine learning that may go by the name artificial intelligence (AI) – are used to make inferences about learning, it is not enough to consider only the elicitation of bits of evidence. The bits may look enticing and seem applicable to a given area of an assessment framework, but in and of themselves, bits of evidence alone do not satisfy a measurement argument.

As described in the Introduction chapter of this report, the interpretation vertex of the "assessment triangle" is actually two things: 1) defensible *elicitation* of bits of evidence; and 2) defensible *accumulation* of this evidence to make an inference. Exactly how all the pieces together add up to satisfy measurement claims and make the intended inferences is key. This accumulation of the bits to make inferences is called *aggregation* of evidence. Both elicitation of evidence and aggregation of that evidence to make claims must be defensible, including showing accuracy and precision of the metrics involved and ruling out alternative hypotheses – as well as verifying that the assessment is fair and equitable for sub-populations (as discussed in Chapter 11 of this report). Before reporting, both the elicitation and aggregation of evidence should be transparent, justified and warranted.

There must also be additional care paid to ensuring that features used in a predictive model are rigorously supported by educational frameworks. Feature engineering in educational assessment contexts should seek to provide meaningful data across levels of student skill rather than focus only on high performing students.

Measurement requires a chain of evidentiary reasoning supported by principled design

When technology is used for the purpose of educational measurement, often a set of questions are asked to guide the development of valid assessment instruments – such as in Evidence-Centred Design (see the Introduction and Chapter 6 of this report for a more detailed description). These questions are phrased differently in different contexts but essentially boil down to the following:

- *What can we do?* In other words, what do technology affordances allow to collect evidence and what "bits" of evidence will be elicited?
- *What do we want to report from the evidence?* In other words, how will we aggregate or accumulate evidence and what are the larger inferences and claims that the "bits" of evidence go together to form so that results can be meaningfully reported?
- *What will be needed to make the connection between the bits of evidence elicited and what we want to aggregate and claim?* In other words, what is needed to make the connection between the first two bullets?

A clear path based on evidentiary reasoning needs to be established between the goals and objectives of measurement (e.g. an assessment framework) and how the evidence is used to make a claim (e.g. the reporting of assessment results). In this way, stakeholders in education are empowered to use and value the results. It is also important to be able to use data to build a solid validity argument, for example verifying that conclusions about students are consistent across tasks that are designed to measure the same set of skills and evaluating that the assessment measures these skills well across the ability distribution and across sub-populations.

However, in many applications of LA based on data-mining approaches, the three principled design questions above are often not asked from the outset. A common feature of these approaches is the idea of discovering results from the data (i.e. generative or exploratory approaches) as compared to building

from theory (i.e. confirmatory). No argument is stated *a priori* regarding the aggregation of evidence but one is determined afterwards based on what is interesting in the evidence – for example, based on patterns in the data that might provide insights into students' cognitive processes in problem solving (Zhai et al., 2020[5]). This type of data exploration can provide highly interesting insights for research but might not be the best approach for making claims about what students can or cannot do. For making a claim, the analytics must be based on a clearly transparent and defensible argument using data.

Prediction is not the same as measurement

Learning analytics often aims to fit the best set of clusters, networks or other structures to data via machine learning, then defend the results based on having chosen the best fitting among various structures. Results are often considered acceptable if the predictions based on these indices are better than other models that were fit and more accurate than random. Patterns over a set of interactions may be possible to consider in measurement models, as discussed in Chapter 8 of this report, but treating predictions for such complex data as if they were valid measurements of latent traits can be problematic as the validity argument is not sufficiently defined nor is the validity evidence accumulated.

Producing a "new" metric just because the market asks might not be best way forward

In policy discussions around large-scale assessments there is often an urgency to assess new constructs and generate new insights, because there are many unanswered questions about students' skills and there is often a sense of *déjà vu* when new reports come out. There are also expectations that applying ML approaches to big data from technology-enhanced tasks will fill all of our existing information gaps. The "market demand" is undoubtedly there but this demand should not be the only driver of decisions in assessment design. What we need is better metrics not just new metrics. It is important to be responsive to policy makers when defining what we should assess and what insights on learners we need, but these new insights must be accompanied by a solid evidence base built on a carefully constructed validity argument. Otherwise, there is the risk that relevant findings are quickly discounted by measurement experts because they are not sufficiently validated and do not meet the "Levy challenges" described earlier.

Some things educational measurement can learn from learning analytics

Embrace the value of naturalistic tasks

Perhaps the most important thing that educational measurement needs to learn from researchers in learning analytics is the use of the naturalistic task. If policy makers want to take advantage of the many affordances of information technology discussed throughout this report, then the measurement field must be able to support the collection and aggregation of evidence from more authentic tasks and more complex activities. In a departure from the standard, multiple-choice assessments of the 20th century, naturalistic tasks are becoming more prominent in educational measurement. Naturalistic tasks can cover a broad array of task designs. Some examples come from science tasks, but many other disciplines also have created naturalistic tasks such navigating through a park in collaborative problem solving or reading with a purpose in mind to gather information for a presentation. In science and many other others, examples include simulations such as students interacting with lab equipment. These tasks are often defined by going through a natural development process where developers create the task similar to a classroom lesson that is guided by content standards (Scalise and Clarke-Midura, 2018[6]) and are further defined by realistic science experiences for students. Embedding small goals of a similar nature, such as assessing inquiry skills, can be done within and across several tasks with as much standardisation as possible in order to use ML/AI techniques to accumulate evidence on student learning (see again Chapter 8 of this report for an example of such an approach).

An important point here is that employing the naturalistic task isn't only about improving the precision and accuracy of the metrics. Measurement scholars will often want to discard complexity in innovation if metrics from simpler item types and tasks are likely to give the same measured result – or one that is similar enough to draw the same or similar inferences. Historically, for instance, in many contexts it has been shown that selected responses can measure some constructs as well as constructed response formats. Yet having students construct unique responses and evaluating those responses with rubrics or a scoring engine is nonetheless a central part of many assessments. Elsewhere in this publication (Chapters 2-5 and 7, particularly), the argument is made that using naturalistic tasks and gathering information on processes improves our evidence base for complex constructs such as collaborative problem solving, when essentially those constructs are largely defined as processes (e.g. collaboration is a process). By contrast, if only response data is considered and no opportunity is given to engage in such a process in an authentic way, the validity argument is certainly affected.

Why else? One reason is the importance of the "signifying" role of assessments. As discussed in the Introduction chapter of this report, educational research has established that teachers, students and local and national policy makers take their cues about the goals for instruction and learning from the types of tasks found on state, national and international assessments. Therefore, what is assessed in areas such as science, mathematics, literacy, problem solving, collaboration and critical thinking, and how those constructs are assessed, often will end up being the focus of instruction. In this role of signifying, it is hence critical that assessments represent the forms of knowledge and competency and the kinds of learning experiences we want to emphasise in classrooms. If students are expected to achieve the complex, multidimensional proficiencies needed for the worlds of today and tomorrow, they should be able to demonstrate their proficiency doing so. This requires moving away from *measuring what is easy* to *measuring what matters*.

Engagement and the student experience are also important considerations. Embedding agency and relevancy in an assessment activity is likely to increase students' engagement and thus the likelihood of observing what students can do at the best of their capacity. If we really want to describe what students know and can do, then test effort is an extremely important assessment argument.

However, we caution assessment developers not to get stuck in the "cluster buster"; not everything has to be as chunky as a long task, which may take a lot of student time and introduce a lot of unique variance that is construct-irrelevant. Educational assessments (at least for now) may need to include a mix of newer and older item and task types and investigate how the evidence produced by different types of task formats and experiences triangulate. Hard-to-measure constructs are often supported by measuring what is proxy and easy. To support interpretations, combining data types that are more known and less known is likely to remain important for making defensible claims.

To improve naturalistic assessments and the quality of measurements they can produce, an iterative and collaborative process between content experts, assessment developers, students and teachers is necessary. These ideas closely align with the concept of "Learning Engineering" as well as the Assessment Triangle discussed in the Introduction chapter of this report, emphasising an iterative process of building affordances and optimising learning experiences. Assessment developers must understand both how teachers look to understand student performance on a given task as well as how students engage with the task environment. These discussions help prioritise task modification to encourage positive task interaction styles among students as well as focus data collection in areas of interest and concern to the teacher.

Consider ways to establish a spectrum of comparability for reporting claims

The assessment field is still poised on the precipice of what assessing competencies using naturalistic tasks means. Costs, versioning, assessment platforms and other practical considerations exist, especially in the context of assessments that are intended to be replicable and comparable. Once such investments are made, often the need to handle longitudinal data also will emerge, further complicating what needs to

be in the measurement paradigm. Therefore, measurement standards may need to find ways to allow entry points that are not as difficult to satisfy. This is the classic solution in other fields with large advances in technology affordances. Can there be tiers and spectrums in education? For example, can there be a comparability spectrum across different purposes for the use of assessment evidence?

This might take the form of co-habiting for a time and focusing on different types of claims. For example, established measurement models might be used to build a scale that describes, in a reliable and comparable way, what students are able to achieve in terms of their outputs from a given set of designed tasks; this could be what problems they were able to solve (given enough information of this type is generated across tasks similar enough to elicit the same latent trait). If this can be done, then LA might be used to provide more descriptive diagnostics of strategies and processes that students follow on the tasks to achieve an output. This might be done through a cluster analysis that describes different "types" of problem solvers, for instance. Descriptions of students' work in each different cluster can be potentially very useful for teachers and students and provide tangible illustrations of applied 21st Century competencies.

Another perspective on the comparability spectrum question may be to not rely so heavily on the perfect fit of each item and score category or each scored observation in innovative assessments. Rather, either patterns (such as those discussed in Chapter 8 of this report) or a factor of larger tolerances might be allowed for the individual observations, if conclusions across the observations would be essentially the same. The same might be concluded about testlets and independence of information, by treating the issue as a discount factor on precision – so not overcounting the information when observations are not entirely independent as it is generally the case in extended naturalistic tasks.

It is hard to say what might be found with further research to simplify the intersection of fields. However, it is well known in educational measurement that no assessment of latent variables ever includes single observations. So less time might be spent in research looking at how individual observations vary, and rather researchers might look at if the inferences over the set vary substantially. Then the focus could be on the comparability of claims made across many items or many observations. However, the analytic mechanisms for how to approach these factors remain to be worked out.

What is at stake?

What is at stake for students may be no less than the development of broad transferable skills and knowledge. Some may believe it is possible to build such skills without assessment and without advanced digital technologies – arguing that such skills were necessary to achieve the accomplishments of earlier times before digital technologies were available. But of course, modern digital affordances are not only useful, as described in several other chapters of this report, but they are also today an expectation in the everyday lives of students. In many cases we don't yet understand the extent to which the broader transferable skills and knowledge that are the target constructs of complex assessments can be elicited across different digital tools – for instance, is self-regulated learning the same latent trait when applied in computational thinking contexts as when applied in reading literacy contexts? Will the knowledge-building tools that we might include in digital assessments to elicit evidence of such skills in turn cause them to manifest differently? What is the interaction of these skills with domain knowledge and domain tools? Such domain specifics no doubt impact the use and therefore performance of any such skills in an assessment context. But learning more about the common strategies students might engage as well as the differences in student behaviours that might arise between applications in digital contexts is important. To the extent there are commonalities, it seems key to understand them.

Conclusion

This chapter has discussed some important ideas that different fields approaching educational assessments might consider. However, an important audience for this report is policy makers. One message to policy makers from this chapter is that it is possible to undermine your objectives by overclaiming from emerging assessments. For instance, if a paramount concern in innovative assessment is ensuring equity and fairness as part of the validity argument, some of the issues discussed in prior chapters will likely require a softening of claims when reporting results from innovative assessments. An example is that policy makers should not select the most enticing wording for the "short" description when reporting on a construct if the wording is inaccurate. Even if this might make the innovative assessment seem more marketable, it will be a problem in the end if the "short" description does not match the claims it seems to be making with high quality evidence.

This will run counter both to what some policy makers want and what the market may demand. The exciting potential of technology affordances and rich new data sets may make policy makers want to lean into making strong claims for very new constructs. A suggestion is to be disciplined and wait on reporting strong claims until the needed progress in measurement science has been made – but don't stop developing and implementing such assessments or it won't be possible to make the needed progress. So do pursue innovation goals or you will never get there, but be mindful of your claims in the meantime.

To summarise, important opportunities will be sacrificed even longer by not incorporating new possibilities from new techniques (i.e. LA) even though this may require considerable exploration and grounding in traditional fields (i.e. educational measurement). Alternatively, important evidentiary techniques for measurement may be lost altogether and need to be recovered with much effort in the future if no pathway is created forward into a modern world. So, we argue, crossover between these two worlds is needed now. Wrestling with these topics will be hard and likely provocative since all sides will not be able to proceed as they currently do. Proceeding without change is also undesirable if fields are to meaningfully inform each other and affordances are to be optimised. Growing organically without incorporating field overlaps will likely lead to many missteps for new analytical approaches and will compromise trust in what the interpretation and use of results can mean for education.

Today, it is too soon to say what solutions might emerge at the intersection of the fields discussed here, but approaches such as hybrid models (see Chapter 8 of this report) or co-habiting (see earlier in this chapter) may become sufficiently established to represent a way forward. Regardless, it would seem inevitable that the bridge to a shared future will mean some spectrum of comparability is needed in educational measurement and assessment.

References

Levy, R. (2012), "Psychometric advances, opportunities, and challenges for simulation-based assessment", *Invitational Research Symposium on Technology Enhanced Assessments, K-12 Center at ETS*, https://www.ets.org/Media/Research/pdf/session2-levy-paper-tea2012.pdf. [3]

Scalise, K. (2012), "Using technology to assess hard-to-measure constructs in the ccss and to expand accessibility", *Invitational Research Symposium on Technology Enhanced Assessments, K-12 Center at ETS*, https://www.ets.org/Media/Research/pdf/session1-scalise-paper-2012.pdf (accessed on 14 March 2023). [4]

Scalise, K. and J. Clarke-Midura (2018), "The many faces of scientific inquiry: Effectively measuring what students do and not only what they say", *Journal of Research in Science Teaching*, Vol. 55/10, pp. 1469-1496, https://doi.org/10.1002/tea.21464. [6]

Sclater, N. (2014), *Code of practice for learning analytics: A literature review of the ethical and legal issues*, Jisc, https://www.wojde.org/FileUpload/bs295854/File/07rp_54.pdf (accessed on 9 April 2023). [1]

Wilson, M. and K. Scalise (2016), "Learning analytics: Negotiating the intersection of measurement technology and information technology", in Spector, J. (ed.), *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice and Policy*, Springer, Cham, https://doi.org/10.1007/978-3-319-17727-4_44-1. [2]

Zhai, X. et al. (2020), "From substitution to redefinition: A framework of machine learning-based science assessment", *Journal of Research in Science Teaching*, Vol. 57/9, pp. 1430-1459, https://doi.org/10.1002/tea.21658. [5]