Embracing Imperfection in Learning Analytics

Kirsty Kitto
Connected Intelligence Centre,
University of Technology Sydney
Sydney, NSW
kirsty.kitto@uts.edu.au

Simon Buckingham Shum
Connected Intelligence Centre,
University of Technology Sydney
Sydney, NSW
simon.buckinghamshum@uts.edu.au

Andrew Gibson
Connected Intelligence Centre,
University of Technology Sydney
Sydney, NSW
andrew.gibson@uts.edu.au

ABSTRACT

Learning Analytics (LA) sits at the confluence of many contributing disciplines, which brings the risk of hidden assumptions inherited from those fields. Here, we consider a hidden assumption derived from computer science, namely, that improving computational accuracy in classification is always a worthy goal. We demonstrate that this assumption is unlikely to hold in some important educational contexts, and argue that embracing computational "imperfection" can improve outcomes for those scenarios. Specifically, we show that learner-facing approaches aimed at "learning how to learn" require more holistic validation strategies. We consider what information must be provided in order to reasonably evaluate algorithmic tools in LA, to facilitate transparency and realistic performance comparisons.

CCS CONCEPTS

• **Human-centered computing** → **User models**; • **Computing methodologies**; • **General and reference** → *Validation*;

KEYWORDS

hidden assumptions, performance, validation, pedagogy, accuracy

ACM Reference Format:
Kirsty Kitto, Simon Buckingham Shum, and Andrew Gibson. 2018. Embracing Imperfection in Learning Analytics. In *LAK '18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3170358.3170413

1 INTRODUCTION

In Learning Analytics (LA) we find the confluence of many tributaries, such as the Learning Sciences, Data Mining, Human-Computer Interaction, and Psychology [18, 48]. While this intersection of fields is exciting, it brings with it all the challenges of interdisciplinarity, including modes of communication, and respect for different quality criteria. As was described by Stember [49] for the social sciences, while fields often claim interdisciplinary status, it is rare to see this characteristic realised in practice. The LA community should regularly check if it seems to be 'playing' at interdisciplinarity, or if a genuine dialogue between disciplines is emerging, one that sees the establishment of new common ground in a move towards transdisciplinarity. Such a dialogue requires an ongoing and critical examination of the hidden assumptions that are imported into LA from other fields; are some fields getting an "easier" treatment because they came with tightly defined methodologies? Here, we will call into question the suitability of appropriating validation criteria from computer science for computational models underpinning *student-facing* feedback, for *particular forms of learning*.

We think it timely, with the rising popularity of data science, to call attention to some of the problems that can infiltrate a field like LA if we do not pay careful attention to our underlying assumptions. Sometimes it is all too easy to evaluate a methodology, using a set of field specific criteria, while leaving out the learning. Therefore, the purpose of this paper is to ensure that the field of LA maintains a balance between its contributing fields, in both the creation of its tools and protocols, and in the ways that we evaluate them.

Specifically, we are going to examine the notion of how LA evaluates its computational tools. We will draw attention to what we perceive as an imbalance, where computational validation methods (which are well defined and therefore easier to use in evaluating tools) can lead to valid LA tools and approaches being inappropriately criticised, or tools with little educational merit being lauded as performing well. More holistic ways of summarising validation across all relevant disciplines are required and we will conclude with a proposal for how LA might work towards achieving this.

1.1 What is learning analytics for?

As was succinctly stated by Gašević et al. [18]: "Learning analytics is about learning". The challenge for LA is to establish plausible relationships between models derived from the neatly quantifiable world of digital data, and the complex socio-cognitive world of "learning". Often we see the validation of different LA tools using measures from the computational sciences, such as accuracy, precision, recall, etc. That is, metrics are used which tell us about *the machine's performance* with respect to a standard, not about *how the machine is enhancing learning as part of the whole system.* The former excludes both the human and the broader educational context in which the use of the LA tool is embedded. We are not against improving the performance of our algorithms *per se*, but will argue here that if the balance of evidence (in the field, or a particular research program, or product development) focuses solely on the metrics of computational performance, then there is no *a priori* reason to expect that these performance gains should translate into improved learning outcomes. We will draw attention to an important class of learning contexts where the notion of a correct prediction, or perfect classification, is far more difficult to define and formalise. This creates challenges for traditional computational

approaches to validation; new ways of proceeding will be required. However, we must first start to unpack the underlying goals behind our LA tools before we can attempt to evaluate them. We begin by considering two broad classes of learning.

1.2 What type of learning?

Over the years an ongoing stream of work in LA has focused upon student facing tools that are used directly in a class context (e.g. [4, 7, 20, 28, 33, 38, 51, 52]). It seems possible to identify two broad motivations behind these solutions: are they teaching students curriculum content, or are they trying to help them learn how to learn more effectively? The way we judge the performance of an algorithm must depend upon a clear understanding of its purpose.

1.2.1 Learning content and skills. The challenges teachers often face revolve around helping students to learn arithmetic, spelling, historical facts, geography, algebra, etc. Reports can help teachers to see if their students have acquired the requisite knowledge, and to identify which parts of the cohort need extra help [4]. A number of projects are starting to personalise messages to students according to their performance in key teacher identified tasks [37]. Often this process of knowledge acquisition can be enhanced by drill and practice models, and educational technology has provided a large number of solutions to this end. For example, Intelligent Tutoring Systems (ITS) have been shown to positively affect student learning outcomes when compared to conventional educational experiences [33, 34, 51]. Closely related, we see adaptive learning [23], and recommendation systems [16, 26] becoming available in many online learning environments. Designed well, these types of systems evidently *assist* students by optimising the pathway to mastering a clearly bounded domain with a curriculum and modes of reasoning that can be formally modelled.

In scenarios of this type it is important that we utilise a model of student learning that closely approximates reality, and that the computational approaches we adopt reflect the underlying reality as closely as possible. After all, an ITS that incorrectly fails a student on a task will be confusing and annoying (to say the least), and could potentially teach the student incorrect content and/or skills.

Less clear cases arise as we start to explore more complex scenarios. Consider, for example, an instructor seeking to help her students to communicate more effectively in a public forum. A number of frameworks have been developed to help people think about this problem [36], and manual qualitative analysis is frequently used by researchers to classify student contributions. Increasingly, LA tools are being developed to automate this classification process using Machine Learning (ML) (see e.g. [13, 29]), and it is common to evaluate these methods by considering how well their classifications overlap with that of human annotators (a point to which we will return shortly). Most of these tools are currently used in a research context, but the aim is to use them in our teaching and learning practice. How might we do this? Two possibilities arise:

An instructor or recommendation system might examine the classifications that have been automatically generated with a view to acting upon this information. In this case it is important that the classifications be highly accurate, as the student profiles generated from this process are being used to assist with understanding student progress.

A student could be shown how their behaviour has been classified by the algorithm. At this point, we argue that the utility of a highly accurate classification becomes more difficult to judge: will the student learn more if they are shown a perfect classification of their behaviour, or one whose accuracy they must judge?

1.2.2 Learning to learn. Student facing contexts open up new possibilities for using LA to help people learn how to learn [12]. Here, we will argue that this second case creates new criteria against which the performance of our computational approaches should be judged. We shall return to this concept in Section 5, but first we will consider an analogy from another field as it increasingly found itself needing to incorporate the user into an analytics loop.

2 A CAUTIONARY TALE FROM INFORMATION RETRIEVAL

How is validation carried out in computer science? One common way to judge the worth of a computational approach involves a consideration of various performance metrics that are defined in terms of how often a task is correctly vs incorrectly performed. For example, when using an algorithm to classify some data trace (e.g. whether a student will get a quiz question correct), we consider whether the algorithm correctly predicted a positive result (true positive, tp), correctly specified a negative result (true negative, tn), or got the response wrong, returning a false negative (fn) or false positive (fp). This must be done with respect to a ground truth dataset that contains the actual student responses. Given these preliminary metrics, we can construct more complex ones:

Precision considers how many times a true positive was returned out of all positive responses:

$$Precision = \frac{tp}{tp + fp}. \tag{1}$$

Recall reports upon how many times a true positive was returned out of all that *should* have been returned:

$$Recall = \frac{tp}{tp + fn}. \tag{2}$$

Accuracy is then defined as the proportion of correctly performed tasks out of all possible tasks (N):

$$Accuracy = \frac{tp + tn}{N} = \frac{tp + tn}{tp + tn + fp + fn}. \tag{3}$$

Considering metrics such as these provides us with a number of ways to compare the performance of different algorithms when applied to the same datasets. They also serve as the basis for more complex metrics, such as the ROC curve, R^2, and Root Mean Square Error (RMSE) [21], some combination of which are frequently reported as performance metrics according to an implicit assumption that error should be minimised, and accuracy should be maximised.

There are fields which have already followed this trajectory. One notable example arose in Information Retrieval (IR), a field which enforces very strict requirements that new models and algorithms be evaluated for improved performance over existing baselines, often with reference to precision based metrics. However, in 2006, Turpin and Scholer [50] published an influential paper that called the entire focus upon precision in IR into question. They considered

mean average precision (MAP), which is a performance metric calculated by taking the mean of the precision scores obtained in a search after each relevant document is retrieved, with relevant documents that are not retrieved receiving a precision score of zero. A higher MAP is indicative of better search results over a series of defined queries in a specified dataset. Turpin and Scholer carefully degraded the performance of a search engine in a toy system, to the extent that it exhibited MAP scores between 55% and 95% at a number of different settings. This search engine was then used in a task that required users to find a single document relevant to a topic, with their performance measured by the length of time taken. Turpin and Scholer showed that there was no significant relationship between system effectiveness measured by MAP and performance in the user task over this range of MAP scores. This suggested that a search engine user is highly unlikely to notice even a large change in an algorithm's precision.

Note that this scenario only arises because a human has been introduced into an analytics loop. While the task of finding a relevant document should in theory have been helped by an algorithm with higher precision scores, the user experience and performance was not significantly improved. There are a broad range of papers that have shown comparable effects [1, 9, 10, 25], although a subset of results appear to suggest that higher precision does correlate with higher user satisfaction ratings [43]. The full implications of putting humans in the evaluation loop are still being investigated in IR, but this line of work has spawned a whole subfield of user modelling approaches attempting to diversify the evaluation metrics used in IR by directly incorporating concepts like user satisfaction, diversity and novelty of results [1, 9, 10]. For example, Clarke et al. [10] have presented a framework for evaluation that systematically rewards novelty (i.e. the need to avoid redundancy in search results, where the same document adds nothing to a user's information need) and diversity (i.e. the need to cover ambiguous terms, such as homonyms which have more than one meaning) in search results. Such an approach might prove effective if attempting to reward a LA tool that encouraged students towards new ideas, creativity and diversity from their current approach.

Since the success of information search is easier to evaluate than the more complex and multifaceted task of learning, it is important to ask a related question: are similar results likely to arise in LA?

3 DOES LA HELP LEARNING?

Interestingly, a recent paper from Educational Data Mining (EDM) calls attention to similar concerns about the validation of computational models in the learning sciences. Liu and Koedinger [30] point out that while EDM aims to improve learning outcomes, its

> *emphasis on the 'educational' aspect of educational data mining has been scarce... One reason for this is the inclination of researchers to evaluate EDM research primarily for model fits and predictive accuracy rather than for plausibility, interpretability, and generalizable insights.*

Instead of relying solely on computational validation metrics, which are "difficult or impossible to interpret" [30, p37], Liu and Koedinger make use of a "human in the loop" analytical component to model the underlying cognitive state of the learner and to then understand how adjustments in a ITS can lead to better learning outcomes. They recommend moving away from automated methods, and towards the mapping of digital traces describing student activity onto interpretable constructs of interest (e.g. Knowledge Components, and the Q-matrix in ITS) which facilitate actionable analytics. Liu and Koedinger [30] achieve balance by demonstrating improved student learning outcomes (using pre and post-tests) *in addition to* their reporting of an improved RMSE value compared to an earlier ITS.

3.1 What is an improvement?

Despite the undeniable successes that have come with computational approaches, it is essential that LA practitioners recognise just how complex the domain of education is when framing our understanding of what they entail. What precisely do we mean by learning? And how can we judge the worth of our algorithms within this understanding? An example will help to illustrate the difficulty associated with asking these questions. Consider Kovanović et al. [29], which demonstrates an automated classification of the "cognitive presence" construct in online discussion fora. This approach achieves an accuracy of 70.3% using a Random Forest model with 205 features and applying SMOTE sampling to correct for unbalanced data across 5 categories of learner event (triggering, exploration, integration, resolution and other, which implies a baseline performance of 20%). Would the result of this paper be *more* convincing if it reported a higher accuracy? In a purely computational field, it would be quite reasonable, and publishable, for another researcher to aim to exceed these performance metrics, but we would like to question the merits of this course of action for LA. That is, we should not follow a path just because it is well understood; we should be asking if the path will actually lead us in a direction that we need to go. Let us consider some other sound papers that were recently published in LAK and EDM from this lens; how likely are they to result in improved learning outcomes?

Consider for example the very interesting work by Allen et al. [2], which attempts to classify a dataset of individual difference measures, text, and keystroke analytics to match self-reported student affective states (in this case boredom and disengagement). While the authors themselves describe their work as preliminary, we note that the accuracy scores reported in this classification as ranging between 76.5% and 77.3% are not likely to be maintained if applied to a genuinely new dataset (rather than using the leave-one-out-cross-validation approach adopted in this paper). The standard deviations for the variables used in the classifier are large compared to the feature values themselves, and the student self-report process is likely to result in a high variability between subjects. Is an improvement of around 25% over a baseline classification of 50% good enough for stability? We will not know until a replication study is performed. Indeed, another paper from the same LAK conference by Buckingham Shum et al. [8] evaluates the performance of a reflective writing analytics tool across multiple datasets, demonstrating a range in accuracy from 70–80% when using the same parameter settings. At what point do we know that any of these classifiers is accurate enough?

Some of the most computationally advanced approaches arise in Knowledge Tracing (KT) scenarios, which seek to model a student's mastery of some body of knowledge. As these approaches have

been trialled for a long period of time, there is a large collection of baseline work against which new models can be compared, and these improvements are often reported using computational metrics with very minor improvements demonstrated either between a new approach when compared to an existing one, or between algorithms compared in the paper itself [39, 40, 44]. While a hypothesis test is often used to denote statistical significance using a p-value, the general lack of effect size data suggests that there is every chance these results will not prove to be replicable [19]. Indeed, Beck and Xiong [5] discuss the way in which many ITS methods have failed to replicate, and even more interestingly, they show that a class of these models is approaching the limits to accuracy that might plausibly be obtained, *despite remarkably low levels of accuracy*. Even if robust ways of improving these performance metrics can be obtained: will they result in better learning outcomes? The results from Section 2 give us reason to pause. More convincing advances are provided by David et al. [14], and the already discussed Liu and Koedinger [30]. These two papers back up their computational validation with user trials that demonstrate improved learning outcomes for students using the new tool over a comparable baseline. This brings us to the first warning of this paper:

Warning 1. *For some educational scenarios, reporting improvement in algorithmic performance is insufficient as a form of validation.*

Before starting to move towards a specification of what we might consider a sufficient validation strategy (in Section 6), we must first examine some of the different ways in which computational evaluation metrics can fail in educational scenarios.

3.2 Measuring the wrong thing

Perhaps one of the most obvious mistakes that could be made would involve reporting upon a metric that has little to do with the task at hand. It is easy to make the mistake of concentrating development in LA upon a concept that is easy to define and track, but not particularly useful to learning. A common example of this problem is provided by dashboards in Learning Management Systems. Not surprisingly, many educators and learning scientists express scepticism about the relationship between these visualisations and learning [22], concerned that feedback about low level user actions such as number of log ins, videos watched, or documents submitted does not illuminate progress in learning, for either students or educators. This failure to provide LA that actually helps learning arises because of an overemphasis upon valuing what we can measure, instead of measuring what we value — a longstanding concern in educational assessment [53]. To take a more advanced example, consider the writing analytics tool that we have been developing [20], which is able to give automated formative feedback on reflective and analytical forms of academic writing. The tool does not replicate spelling and grammar checkers, although the team could have invested effort in perfecting the associated code to report metrics on how improved different versions were. Such an 'improvement' in accuracy would have distracted developer effort and student attention from a focus on thinking about the higher order rhetorical moves in their writing (the purpose of the tool).

A focus on improving the wrong analytics will contribute nothing to student learning, and yet this is an easy mistake to make. We have arrived at the next warning of this paper:

Warning 2. *Being* able to *report upon a metric does not mean that you* should *use it, either in the tool, or in reporting its worth.*

4 THOUGHT EXPERIMENTS IN PERFECT CLASSIFICATION

Once we have (i) understood what type of learning we are trying to facilitate in our students, and (ii) are sure we are measuring the right thing, we must start to consider (iii) precisely how accurate our student facing LA must be to assist this outcome. Here, we will introduce two *perfect classification* thought experiments. This will help us to clarify why perfect accuracy does not 'solve' the LA design challenge. On the contrary, we will argue that imperfect analytics will sometimes be useful in enhancing student learning.

4.1 Writing Analytics

Returning to the writing analytics example of Gibson et al. [20], this work seeks to make visible those aspects of writing where a student has made a "rhetorical move" commonly found in academic argumentation, or professional reflection. The authors describe the set of moves, the mechanisms for implementing them, and the degree to which an automated classifier matches human judgement, which by computational standards leaves much room for improvement.

It would seem desirable then, to create a system which could identify *all* of the relevant rhetorical moves with perfect accuracy, and then provide feedback for the student about where they occur and when they are missing.

However, there are at least three reasons to be cautious about adopting such an approach. Firstly, *the way students learn to write is not the same as the way experts make sense of writing*. While identifying rhetorical moves may assist in the analysis of writing, students do not necessarily learn to write by stringing together a series of rhetorical moves. There is a large amount of "intellectual infrastructure" that a person must build as they gain expertise, so we would want to first ask questions such as: Can this student understand the concept of rhetorical moves? and; Would it be more effective to design writing improvement activities that give them an opportunity to practice this skill first? A computational approach to validation has nothing to say on this issue.

Secondly, because the machine can read an essay in less than a second and return real time feedback, the student could easily be overwhelmed by too much feedback — e.g. a report might be generated which tells a student 42 things they should do to improve their draft. Rather, we would want to ask what the most important elements to foreground are, at the current stage in a student's journey to becoming an accomplished writer. This is a standard strategy used in teaching writing: an experienced PhD supervisor knows not to provide all their feedback at once (which would be the most accurate solution); they provide it in manageable chunks, tailored to each student.

Finally, it is important to differentiate between *improved outcomes*, such as submitting a good piece of writing, and *learning how to write* in a way that will translate to new contexts. It is entirely conceivable that students might 'correct' their work in the light of the feedback they receive from a LA tool, finishing their essay faster and obtaining higher grades. Would we count this as an analytics success story? Arguably not, particularly if the student had

failed to internalise the reasons for the improvement, and could not translate the skills to new writing without being dependent on the tool. At this point we would have a partial analogy to writers who are dependent on spelling and grammar checkers: they have 'outsourced' this task to the machine.

This thought experiment suggests that even if a machine *could* perfectly identify all of the rhetorical moves in a piece of writing, it does not follow that it *should* show everything to a student to nurture the right kind of learning (i.e. learning to write).

Warning 3. *Feedback should not necessarily be set at the same resolution that the analytics make possible.*

Since writing is an extremely complex form of learner trace to analyse, with infinite expressive nuance and potentially multiple levels of meaning, it is reasonable to assume that in fact, perfect analytics are *in principle* not going to be possible — they will remain a thought experiment. However, imperfect as it is, the writing analytics tool of Gibson et al. [20] is being piloted with students. These trials are leading to largely positive student self-reports and indications that the tool leads to writing process improvements such as more redrafting [47]. While evidence is still required that this tool improves the writing product produced by students, when we are dealing with higher order competencies, imperfection need not be the enemy of the good when it comes to learning analytics, as the following example also illustrates.

4.2 Online forum analytics

The "cognitive presence" classifier discussed in Section 3.1 is currently the best performing classifier for this educational construct [29]. While it achieved an accuracy of 70.3% for the dataset on which it was trained, it is unlikely to perform as well on a different dataset. Even if this classifier demonstrated performance metrics above some threshold for a defined collection of datasets, it is unlikely to maintain this performance across all educational scenarios, especially given the variability in online discussion fora. Should its status therefore be confined to that of a research prototype until it can score 80-90%? Perhaps we need an even better accuracy? Or could it already be deployed with students in some way?

We have a second line of research leading to the development of an "Active Learning Squared (AL^2)" paradigm [28], which makes use of a cognitive presence classifier to scaffold student metacognition. AL^2 is so-called because it seeks to promote both active/self-regulated student learning [54] and active machine learning (i.e. aiming to reduce the amount of time required to create a labelled corpus that will be used to train the classifier e.g. [46]). AL^2 can only make use of an imperfect classifier because the methodology is coupled with a tight learning design that requires students to understand *why* they are participating in the activity [27]. In the trials that have been run to date [28] this activity is used to help students learn (i) how machine learning classifies text, and (ii) what their profile of behaviour in an online discussion forum looks like.

What is the point of this activity from the perspective of the student? In this scenario the student is encouraged to reflect upon how the algorithm is being used to classify their behaviour, and to challenge classifications that they think are wrong. This learning activity is designed to help students to open up the black box of machine learning [6, 35] and to question the way in which it may be inappropriately used. Therefore, the AL^2 tool aims to *increase* the cognitive load of a student, slowing down their heuristic thinking and helping them to drop into a more thoughtful or reflective mode [24]. The aim of this scenario is therefore not to teach a student how to classify text (although they might also learn how to do this). At its core AL^2 aims to improve students' understanding of how machine learning works (and specifically that it can be wrong) while teaching them about an educational construct that is likely to help with their participation in a common educational scenario (i.e. communicating in a discussion forum). Note that with this understanding, the learning task (building data literacy) is different from the activity that is being carried out by the student (correcting the classifications). We demonstrated that when embedded in good learning design, students can engage productively with this imperfect analytics tool to reflect more deeply about their behaviour. However, we do not yet have conclusive evidence that in this pedagogical setting, student outcomes are improved *because* the classifier is imperfect.

We can, however, reflect on this claim via a second thought experiment. It is plausible that a hypothetical student could move through the learning activity with little underlying motivation to question how their behaviour has been classified — the machine is always correct after all, and they are merely 'rubber-stamping' its decisions. However, as D'Mello and Graesser [15] demonstrate, it is when the student experiences dissonance because the analytics fail to match their expectations that they are likely to reflect on why they think the machine is wrong. We believe that this form of critical questioning is more likely to happen if the student has been given an underlying reason to be a little distrustful of the classifier. But how imperfect can the classifier be? Is there an optimal level of misclassification for student learning in scenarios like this? This line of questioning is yet to be pursued in the LA community.

4.3 Using imperfect machine learning now

It is worth pausing at this point to highlight the new avenue that AL^2 has opened up. This approach arose from asking an important question: how accurate does Machine Learning (ML) need to be before we can safely use it? There is no immediate reason to believe that classifiers will maintain performance across all educational scenarios. Education is a field which contains an enormous number of potential features, hard to capture contextual influences and other confounding variables. If we insist on prioritising computational approaches to validation then a classifier with better precision, recall etc. should be preferred. But how much data is enough to train a sufficiently accurate classifier? And how will we know that we have trained the classifier over all relevant classes of student behaviour? It is usually assumed to be important that classifiers be accurate, as otherwise a student will be subjected to inappropriate interventions. However, such a position leaves us in a dilemma; are we to wait until perfect accuracy is achieved?

In Section 4.2 it was the bricolage nature of LA itself that suggested an alternative approach, one based upon an educational paradigm. Instead of waiting for an unrealistic error free classifier that we can use without fear, the problem has been reframed by reconsidering the underlying purpose of classifiers in LA. Are they being used to teach skills and knowledge? Or, can they be used

to encourage our students in learning how to learn? In this paper we have so far seen imperfection used in examples of learning to think and write critically and reflectively, as well as learning how to contribute critically and constructively to a discussion, but many more opportunities present. Learning to learn requires only that our algorithms encourage students to think deeply as a result of some activity, not that they be perfectly accurate. Indeed, we might ask what most helps a student to learn; a perfectly accurate classification? Or a provocation that they might challenge? We see the AL^2 paradigm as an early example of following this questioning through to its logical conclusion, bringing us to our final warning:

Warning 4. *Overemphasising computational accuracy is likely to delay the adoption of LA tools that could already be used productively.*

5 LEARNING HOW TO LEARN IN AUTHENTIC CONTEXTS

Section 1.2 recognised the different types of learning that LA needs to support, and Section 3 noted that there is a tendency for LA tools to be evaluated in computational terms, rather than with reference to improved learning outcomes. We then briefly considered two examples as a way of introducing the idea that perfect machine classification of student behaviour may be unattainable, and possibly even educationally undesirable. In this section we characterise those examples as falling into a broader class of contexts — those in which "learning how to learn" is an explicit learning objective that arises alongside another one: that of a rapidly evolving future where "equipping students with knowledge, skills, and dispositions that prepare them for lifelong learning, in a complex and uncertain world" [7, p6] is increasingly important.

Creativity, critical thinking, agency, curiosity, and an ability to tolerate uncertainty are increasingly emphasised in curricula around the world. This shift also entails the use of more authentic learning contexts and assessments that are designed to create conditions where these qualities can differentiate learners. Thus, the emphasis is shifting towards 'wilder' learning environments, which are more open and difficult to control than a school or university classroom. These situations provide LA with more open-ended challenges, and a wide range of complex characteristics. They include:

- **Embodied, skilled performance:** Scenarios in which an important part of the learning experience involves material that is not 'online' but physically embodied (e.g. inspecting a forest; a nursing ward; conducting a social services risk assessment). This embodiment makes it both far more complex, if not impossible, to tightly control what will happen, as well as making outcomes far harder to digitally monitor.
- **Transformed perspective:** Assessments where we focus upon the sense that a learner can make of their experience, or a shift in worldview, which by definition is not accessible to the machine, but to which a machine might have partial access (e.g. a reflective journal on a work placement).
- **No correct solution:** Genuinely complex 'wicked' problems that have no correct solution, only better or worse interventions, which makes definitions of 'mastery' hard, if not impossible to formalise (e.g. a group project to devise a homelessness strategy that is acceptable to the homeless, the police and residents).
- **Socially and psychologically complex performance:** Scenarios where the focus of assessment is emergent in nature, a function of many drivers that result in unpredictable and/or unique outcomes, often because social interaction is central to the process (e.g. the quality of a therapeutic session, or of a conflict resolution process).

Analytics based approaches for tackling some of the above are emerging [7], but returning to our thought experiments, we might expect that such analytics will *in principle have a high degree of imperfection*: it is likely to be impossible for computers to sense all relevant environmental variables, and to formally model the relationships between those variables and learners' psycho-social states. Rather, teaching in such situations requires educators to design the conditions in which learners cultivate the disposition to engage with feedback, and by extension LA systems, in mindful ways. This in turn requires that designers of LA infrastructure create software with the affordances to encourage such active engagement, and learning design patterns for the coherent integration of such tools into the learning experience. We elaborate on this next.

6 EMBRACING IMPERFECTION

How might we move beyond an overemphasis upon trying to eliminate imperfection in the algorithms that LA utilises? Innovations such as this form part of an essential transition for new interdisciplinary fields. Rather than adopting tools and methods from other fields and applying them to a new context (a scenario that is normally understood as cross-disciplinary) it is essential that interdisciplinary fields transition to idea generation as they mature, a concept that was recently discussed in a similarly bricolage field, Human-Computer Interaction, where Liu et al. [31] argued that interdisciplinary fields must start to create their own *motor themes*, or dominant paradigms, in order to achieve focussed research direction. We propose that *embracing imperfection* could be a new mode of conceptual understanding that comes from a genuine intersection and extension of LAs contributing disciplines [49].

In this section we will discuss key concepts that we believe will help LA to embrace imperfection in student facing solutions. This will facilitate the discussion in Section 6.3, where we operationalise the concepts in this paper with a suggestion for a more holistic evaluation of LA tools that could be used to enhance transparency and cross comparison in scenarios where this becomes necessary.

6.1 Aligning LA with Learning Design

Firstly, referring to the discussion of Section 4.2 we see that the imperfect ML applied in the AL^2 paradigm was used only with careful learning design (LD). Kitto et al. [28] have recently demonstrated that a tight integration of LA with LD appears to help generate more reflective students for one set of scenarios. In that case, a set of trials that required students to write reflective blogs about their participation in the online community carefully increased the coupling of LA to LD, with an apparent improvement in student learning outcomes. However, this improvement is hard to measure. How can we judge the validity of the approach?

Similarly, referring back to Section 4.1, more recent work using the AWA tool uses LD to scaffold an entire writing improvement activity [47]. Students are guided through a series of tasks, such

as understanding the instructor's rubric, improving a sample text, reviewing exemplar improvements, self-assessing their work, and reflecting on the quality of the automated feedback. This LA toolset is designed in a modular fashion to support the learning design used by an instructor, who can select the task components to be included, and personalise the feedback experience for different students.

Both of these scenarios align with the notion of process analytics that was first discussed by Lockyer et al. [32], in which students are guided towards critical reflection about their behaviour in performing activities. They are given many opportunities to formatively engage with and reflect upon the automated feedback that they receive along the way, building to the point where they are formally assessed on the understanding that they have gained in the process.

Note that in both of the above examples, the students are explicitly alerted that the analytics could be incorrect in some way, and this is a feature that they should work with:

> *. . . students are given a very brief one page tutorial about what cognitive presence construct is, and then encouraged to enter into an activity where a classifier scaffolds them during their analyse phase. A screen shows them how specific posts they have made in their learning community have been classified, and instructs them to think about this classification and to correct it if they think it is wrong. They are also encouraged to record reasons for this reclassification, and to highlight features in the post they think are indicative of their new classification.* [28, p157]

and

> *. . . students should be encouraged to argue with the machine when they disagree with the feedback. Assuming there is an acceptable signal-to-noise ratio, this is exactly the higher level of discourse that we want to provoke. Academics have often proposed to us that they could envisage productive collaborative activities in which pairs of students use their AWA reports as a springboard for discussion with each other.* [8, p78]

Thus, both scenarios encourage students to critically reflect upon what the LA says about them, and to challenge those analytics if they think that they are incorrect. It is essential that the field of LA operationalise the way in which it discusses this linking of analytics with LD, and test the ability of our tools to facilitate this process.

To summarise, it is the careful use of LD that both mitigates for, and indeed takes advantage of, imperfect analytics. The gap between the learner's model and the machine's model (of the world and the student's learning), creates a 'teachable moment', a dissonance that requires resolution. The machine's limitations are thus compensated for by the human's intelligence: *the burden is placed on the learner to make sense of the analytics, in a process that has been designed to advance their learning*. This approach is an example of how LA can combine AI with IA: Intelligence Augmentation, as argued for by Engelbart in 1962 [17]. IA advocates that new software tools for intellectual work must be co-evolved with new work practices and human training (for another example in the context of ITS, see Baker's argument [3]). Once the analytics is embedded in an appropriate learning design we can see that its purpose is to provide enough scaffolding to "start a conversation" between the student and the analytics-driven feedback, or between peers. Here we see formative feedback coming to the fore in scenarios that aim to encourage students to think more deeply about their own behaviours and perhaps to modify them. Concepts like computational accuracy should be reframed as important only *to the degree that they facilitate this process*.

We argue, therefore, that imperfections in our computational approaches are not only intrinsic to more complex, authentic learning scenarios, but are in fact *a feature to be exploited in well designed learning activities*. Instead of eliminating imperfection from our models, we can use it as a strategic asset that can be embraced in the pursuit of analytics-informed approaches to the kinds of learning discussed in Section 5. Taking advantage of this opportunity will require a shift in mindset, away from scenarios where analytics invisibly control what learners can see and do, and towards scenarios where learners are provided with tools that help them take responsibility for their learning, and reflect critically on automated feedback. We will also require new ways to gauge the extent to which our tools facilitate this process (a point to which we shall return in Section 6.3).

6.2 Designing for mindful student engagement with automated feedback

In a widely cited paper from 1991, Salomon et al. [42] reflect on the relationships that learners can have with educational technologies, and hence, how researchers should frame the task of evaluating them. Should the distributed intelligence of the whole system's performance (humans + technology) be the output measure? (It often far exceeds humans working alone.) Or, should we also be concerned with the effects on human performance when stripped of the technology? This paper seems as topical now as it was 27 years ago in a different digital era, and merits a far deeper examination than space permits here.

We will constrain ourselves to drawing attention to a specific concept that they introduce, which is particularly relevant to the present discussion, namely the concept of *mindful engagement* with technology. Salomon et al. [42] are concerned that students move beyond *mindless* use of potentially powerful cognitive tools, and instead employ "nonautomatic, effortful, and thus metacognitively guided processes" [p4]. This is precisely the role that we have been arguing that "imperfect analytics" can help to facilitate, and the kinds of automated feedback that they can give. The fact that the learner is required to work harder to assess what they are being presented with is *a feature, not a bug*.

6.3 But then how should we evaluate success?

This paper has made the argument that computational metrics can be misleading in a wide variety of ways. We are of course not arguing that they should be discarded, but rather that they must be considered within the larger context of learning, especially when being used to evaluate student facing LA that has a "learning to learn" focus. What then would this larger context entail? Many different groups have proposed frameworks for LA that consider the way in which we must complete the loop to return analytics to the user (see e.g. [11, 41, 45]). However, we are yet to see considerations of this process that explicitly call attention to the interaction

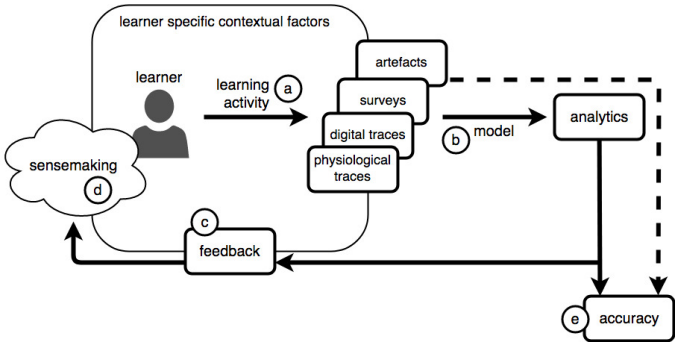

Figure 1: Five forms of validation in the analytics cycle.

between accuracy and opportunities for more complex metacognition in student facing solutions. In Figure 1 we give a general representation of key aspects of these frameworks. Taking it as a reference point we can see that the computational approach to validating LA emphasises the path traced out by the dashed arrow, whereby a signal produced by the learner is modelled and analysed, and the outputs of this process are then tested for accuracy, precision, etc. according to a reference standard, which in the best cases is an open dataset and associated metrics.

We would like to draw attention to some points in the LA cycle (denoted by letters (a)–(d) in Figure 1) that this paper has shown tend to be less well considered, but which must nevertheless form a part of any systemic LA evaluation strategy. Specifically, there are four additional points where we think *all* student-facing LA tools should be held accountable when evaluating them:

(a) **What is the learning activity in which this LA tool should be used?** This could be specified using a well articulated learning design which describes the learning activity that a learner engages in. What are the underlying motivations of the activity? What data objects are produced? And which ones are ignored?
(b) **What model is used to link these low level data traces to the analytics?** Models are often a hidden assumption behind a LA cycle, but they should be both specified and justified (especially if there is no well established educational, social or psychological model driving the analytics cycle).
(c) **What form of feedback is provided to learners?** Has it been designed according to a well understood process? Has it been evaluated by end users? What metrics can be used here to judge the worth of the tool? Note that a user evaluation of LA tools (while important) is not adequate for a complete understanding of its value.
(d) **In what ways does the feedback contribute to any form of learning gain?** Our students need to be able to interpret and make sense of the reports that have been produced, reflecting upon what it means to them, and whether they should change their participation in the learning activity. Few studies attempt to do this, and they tend to belong to the content and skills type learning scenarios discussed in Section 1.2. We have even fewer ways of responding to this question when we consider scenarios where our students are learning how to learn.

We propose that a standard and holistic approach to validating student facing learning analytics should be required to report upon these 4 components of the learning analytics cycle *in addition to* reporting upon relevant computational performance metrics such as accuracy. The LA community could move forwards in this important area if it were to adopt a standardised template for reporting upon all components of the LA cycle. Such a template would push groups beyond implicit assumptions that e.g. computational metrics are sufficient as a quality indicator, and give us a way of comparing the maturity of different approaches and tools. Some of the slots will be empty for some approaches. This is not necessarily a problem at the early stages of a tool's development, but if a group consistently fails to provide information on a specific point then this should be seen as a problem for the LA tool, detracting from its usefulness.

What would this standardised approach look like? We will illustrate an early attempt at this more comprehensive reporting of LA tool performance using the two specific instances of student facing tools that have been discussed throughout this paper.

6.3.1 *Reflective Writing Analytics, RWA, [20].*

(a) **Learning Design:** The purpose of this tool is to teach students how to produce more reflective writing. The software has been designed to accommodate the learning design of individual subjects, with each subject also drawing on the theoretical model which informed the design of the analytics.
(b) **Model:** The model was developed from educational theories of reflection and reflective writing, and is informed by Systemic Functional Linguistics [20].
(c) **Feedback:** Students are presented with annotations based on sentence and sub-sentence level feature alignment with the model. They are also provided with textual feedback assisting with the interpretations of the annotations.
(d) **Sensemaking/Gain:** The students engage in a sensemaking process that connects knowledge of the theoretical model as presented in the subject learning resources (including a rubric) with the annotations they view in the text. Annotations are an affirmation that they are 'on the right track'. Where annotations are limited, the textual feedback draws student attention to missing elements, and the student is required to return to the subject resources to address the deficiencies in their writing.
(e) **Accuracy:** Buckingham Shum et al. [8] document conventional classification metrics for an early version of the reflective writing parser (Precision 0.509; Recall 0.623; Accuracy 0.799; F1 0.56). The complexities of establishing a human gold standard for a construct such as 'reflection', as well as the limitations of the parser are identified as contributing to the relatively poor results. However, while these are not strong metrics by conventional standards, the tool is being used in class contexts, since the limitations of machine intelligence are made up for by reminding students that they have the agency to reflect critically on the feedback, and the way in which the tool is aligned with the curriculum materials guides them to reflect on the extent to which they have addressed the subject requirements.

6.3.2 *Active Learning Squared, AL^2, [28].*

(a) **Learning Design:** The purpose of this tool is twofold: it should teach students data literacy; and help them to learn about the

educational framework against which their behaviour has been classified. A LD pattern has been developed alongside the AL^2 tool, but its utility in facilitating student learning has yet to be tested beyond a preliminary pilot stage.

(b) **Model:** This approach relies upon a dual process model of cognition [24]. Thus, it hypothesises that student facing LA can be designed to encourage students to drop into 'slow' and deep reasoning processes from 'fast' heuristic approaches. To date this process has been coupled with the Community of Inquiry model that drives the "cognitive presence" construct used in the classifier [29], but other models could be used as appropriate classifiers are developed.

(c) **Feedback:** Automated classifications from ML are appended to student comments and presented in a new display.

(d) **Sensemaking/Gain:** Students engage in AL^2 via a web based dashboard that first presents them with a one page explanation of the educational schema that the classifier is applying to their text, and a warning that it is not completely accurate. They then navigate through a series of pages where each post that they have made is presented, along with the forum context in which it appears. The interface allows the student to (i) change the classification of their post, (ii), highlight components of the post that they feel are indicative of the classification they have chosen, (iii) leave a comment about why they chose that classification. Thus, this dashboard has been very carefully designed to encourage sensemaking. User trials of this dashboard have been preliminary pilot studies, and so evaluations of how successful the tool is in developing student data literacy are yet to be performed. Trials are planned for 2018.

(e) **Accuracy:** The accuracy of the classifier used in the initial pilot trials was very low. A simple Naive Bayes classifier was implemented (accuracy 30.2%) rather than the state of the art solution as this was not available at the time of the trials. The performance of that state of the art solution when trained on its original data set (without the subset of Coh-Metrix features due to the closed nature of that code base) on the same data was 47.3% when applied to the data set with the SMOTE sampling turned off during the training process, with performance dropping to 30.5% when the classifier was trained using SMOTE sampling. This generally low performance points to potential overfitting of the best performing classifier to its training dataset.

6.3.3 A comparison. More work obviously remains to be done, but within this reporting format we can start to see some similarities and differences between the two approaches. For example, RWA has been designed to be adapted to the specifics of a course by any teacher who would like to improve the reflective writing capabilities of their students. This adoption will require careful examination of the tool and the creation of well thought out learning designs that teachers can use 'off the shelf' in a manner similar to the new work completed by Shibani et al. [47]. In contrast, AL^2 has a specific learning design already, which could be extended and used with many classifiers of student behaviour. However, use cases for the tool are reasonably restrictive and cannot be adapted with the same flexibility as RWA. Both approaches have well defined models driving the LA, a feature that is likely to be increasingly necessary, as we discover that low level clickstream data is not generally amenable for model free extraction into educationally relevant reports. Both approaches have well developed feedback capabilities, and a strong emphasis upon student agency for sensemaking. Neither have been well tested in terms of the learning gains that they generate; an area that is now a high priority for both research and development programs. Note that neither approach rests solely on the accuracy of the computational methods that the tools use. Had the evaluation strategy considered only accuracy, then neither tool would have been deployed with students.

7 CONCLUSIONS

We have argued that inappropriate outcomes are likely when computational approaches are used to evaluate a specific class of student facing LA solutions (i.e. those aiming to help our students learn how to learn).

Rather than importing established computational paradigms for the validation of our tools, we contend that there is significant opportunity for LA to question those assumptions, and work on developing new validation criteria that emphasise learning outcomes. A preliminary proposal for operationalising these ideas has been exemplified with the Reflective Writing and Active Learning Squared analytics examples. However, while designing, deploying and evaluating those two examples has served to ground our thinking, we have identified a pattern in the 'signals' we are seeing from thought leaders in related fields such as: Engelbart (Intelligence Augmentation); Baker and Koedinger (Intelligent Tutoring Systems); Salomon (Educational Technology); Turpin and Scholer (Information Retrieval). Each of these contributions point to the complexity of evaluation when a human "is in the loop", and together with this paper, provide mounting evidence that embracing imperfection is a rich and fertile avenue of research to pursue.

To conclude, our hope is that this paper serves as both a cautionary tale, and a provocation to open up a new direction in which the LA community might choose to travel. As an emerging field we must constantly check the assumptions embedded in the worldviews and technologies of our diverse constituent disciplines. Learning Analytics has the chance to mature with new approaches that grow from a genuine dialogue and mutual adaptation of its contributing fields. Can the field earn a truly transdisciplinary status?

8 ACKNOWLEDGEMENTS

Our thanks to Abelardo Pardo for early discussions that began in 2016. We gratefully acknowledge support from the Australian Government Office for Learning and Teaching (OLT). The views in this paper do not necessarily reflect the views of the OLT.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*. ACM, 5–14.

[2] Laura K Allen, Caitlin Mills, Matthew E Jacovina, Scott Crossley, Sidney D'mello, and Danielle S McNamara. 2016. Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 114–123.

[3] Ryan S Baker. 2016. Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 600–614.

[4] Aneesha Bakharia, Linda Corrin, Paula de Barba, Gregor Kennedy, Dragan Gašević, Raoul Mulder, David Williams, Shane Dawson, and Lori Lockyer. 2016. A

conceptual framework linking learning design with learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 329–338.

[5] Joseph Beck and Xiaolu Xiong. 2013. Limits to accuracy: how well can we do at student modeling?. In *Educational Data Mining 2013*.

[6] Simon Buckingham Shum. 2016. Algorithmic Accountability for Learning Analytics. (2016). Seminar, University College London, April 2016. Retrieved from http://bit.ly/aala2016.

[7] Simon Buckingham Shum and Ruth Deakin Crick. 2016. Learning Analytics for 21st Century Competencies. *Journal of Learning Analytics* 3, 2 (2016), 6–21.

[8] Simon Buckingham Shum, Ágnes Sándor, Rosalie Goldsmith, Randall Bass, and Mindy McWilliams. 2017. Towards Reflective Writing Analytics: Rationale, Methodology and Preliminary Results. *Journal of Learning Analytics* 4, 1 (2017), 58–84. https://doi.org/10.18608/jla.2017.41.5

[9] Charles Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An effectiveness measure for ambiguous and underspecified queries. *Advances in Information Retrieval Theory* (2009), 188–199.

[10] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 659–666.

[11] Doug Clow. 2012. The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd international conference on learning analytics and knowledge*. ACM, 134–138.

[12] Ruth Deakin Crick, Cristina Stringher, and Kai Ren. 2014. *Learning to learn: International perspectives from theory and practice*. Routledge.

[13] Sebastian Cross, Zak Waters, Kirsty Kitto, and Guido Zuccon. 2017. Classifying help seeking behaviour in online communities. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 419–423.

[14] Yossi Ben David, Avi Segal, and Ya'akov Kobi Gal. 2016. Sequencing educational content in classrooms using Bayesian knowledge tracing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 354–363.

[15] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[16] Hendrik Drachsler, Katrien Verbert, Olga C Santos, and Nikos Manouselis. 2015. Panorama of recommender systems to support learning. In *Recommender systems handbook*. Springer, 421–451.

[17] Douglas C. Engelbart. 1963. Conceptual Framework for the Augmentation of Man's Intellect. In *Vistas in Information Handling*. Spartan Books, 1–29.

[18] Dragan Gašević, Shane Dawson, and George Siemens. 2015. Let's not forget: Learning analytics are about learning. *TechTrends* 59, 1 (2015), 64–71.

[19] Andrew Gelman and Eric Loken. 2014. The Statistical Crisis in Science Data-dependent analysis. *American Scientist* 102, 6 (2014), 460.

[20] Andrew Gibson, Adam Aitken, Ágnes Sándor, Simon Buckingham Shum, Cherie Tsingos-Lucas, and Simon Knight. 2017. Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 153–162.

[21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning with applications in R*. Vol. 112. Springer.

[22] Ioana Jivet, Maren Scheffel, Hendrik Drachsler, and Marcus Specht. 2018. License to evaluate: Preparing learning analytics dashboards for the educational practice. In *Proceedings of the International Conference on Learning Analytics and Knowledge, Sydney, Australia, March 2018 (LAKâĂŹ18)*. In Press.

[23] Dale Johnson. 2017. Opening the Black Box of Adaptivity. *Educause Review* (2017). Available at http://er.educause.edu/blogs/2017/6/opening-the-black-box-of-adaptivity.

[24] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[25] Diane Kelly and Cassidy R. Sugimoto. 2013. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the Association for Information Science and Technology* 64, 4 (2013), 745–770.

[26] Hassan Khosravi, Kendra Cooper, and Kirsty Kitto. 2017. RiPLE: Recommendation in Peer-Learning Environments Based on Knowledge Gaps and Interests. *Educational Data Mining* 9, 1 (2017), 42–67.

[27] Kirsty Kitto, Mandy Lupton, Kate Davis, and Zak Waters. 2016. Incorporating student-facing learning analytics into pedagogical practice. In *Show Me The Learning. Proceedings ASCILITE 2016*. 338–347.

[28] Kirsty Kitto, Mandy Lupton, Kate Davis, and Zak Waters. 2017. Designing for student-facing learning analytics. *Australian Journal of Educational Technology* 33, 5 (2017), 152–168.

[29] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards automated content analysis of discussion transcripts: a cognitive presence case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 15–24.

[30] Ran Liu and Kenneth R Koedinger. 2017. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *JEDM-Journal of Educational Data Mining* 9, 1 (2017), 25–41.

[31] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3553–3562.

[32] Lori Lockyer, Elizabeth Heathcote, and Shane Dawson. 2013. Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist* 57, 10 (2013), 1439–1459.

[33] Marsha Lovett, Oded Meyer, and Candace Thille. 2008. JIME — The open learning initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education* 2008, 1 (2008).

[34] Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. (2014), 901-918 pages.

[35] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[36] Tim O'Riordan, David E. Millard, and John Schulz. 2016. How should we measure online learning activity? *Research in Learning Technology* 24, 1 (2016), 30088.

[37] Abelardo Pardo, Jelena Jovanović, Shane Dawson, Dragan Gasevic, and Negin Mirriahi. 2017. Using Learning Analytics to Scale the Provision of Personalised Feedback. *British Journal of Educational Technology* (2017). In review.

[38] Abelardo Pardo, Negin Mirriahi, Roberto Martinez-Maldonado, Jelena Jovanović, Shane Dawson, and Dragan Gašević. 2016. Generating actionable predictive models of academic performance. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 474–478.

[39] Zachary Pardos, Yoav Bergner, Daniel Seaton, and David Pritchard. 2013. Adapting bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013*.

[40] Zachary A Pardos and Yanbo Xu. 2016. Improving efficacy attribution in a self-directed learning environment using prior knowledge individualization. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 435–439.

[41] Bart Rienties, Avinash Boroowa, Simon Cross, Chris Kubiak, Kevin Mayles, and Sam Murphy. 2016. Analytics4Action evaluation framework: A review of evidence-based learning analytics interventions at the Open University UK. *Journal of Interactive Media in Education* 2016, 1 (2016).

[42] Gavriel Salomon, David N Perkins, and Tamar Globerson. 1991. Partners in cognition: Extending human intelligence with intelligent technologies. *Educational researcher* 20, 3 (1991), 2–9.

[43] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 555–562.

[44] Michael Sao Pedro, Ryan Baker, and Janice Gobert. 2013. Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *Educational Data Mining 2013*.

[45] Maren Scheffel, Hendrik Drachsler, Slavi Stoyanov, and Marcus Specht. 2014. Quality indicators for learning analytics. *Journal of Educational Technology & Society* 17, 4 (2014), 117.

[46] Burr Settles. 2012. Active Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning). *Long Island, NY: Morgan & Clay Pool* (2012).

[47] Aileen Shibani, Simon Knight, Simon Buckingham Shum, and P. Ryan. 2017. Design and Implementation of a Pedagogic Intervention Using Writing Analytics. In *Proceedings of the 25th International Conference on Computers in Education*.

[48] George Siemens. 2013. Learning analytics: The emergence of a discipline. *American Behavioral Scientist* 57 (2013), 1380–1400. https://doi.org/0002764213498851

[49] Marilyn Stember. 1991. Advancing the social sciences through the interdisciplinary enterprise. *The Social Science Journal* 28, 1 (1991), 1–14.

[50] Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 11–18.

[51] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4 (2011), 197–221.

[52] Katrien Verbert, Erik Duval, Joris Klerkx, Sten Govaerts, and José Luis Santos. 2013. Learning analytics dashboard applications. *American Behavioral Scientist* 57, 10 (2013), 1500–1509.

[53] Gordon Wells and Guy Claxton. 2008. *Learning for life in the 21st century: Sociocultural perspectives on the future of education*. John Wiley & Sons.

[54] Philip H Winne. 2010. Improving measurements of self-regulated learning. *Educational Psychologist* 45, 4 (2010), 267–276.