Towards more replicable content analysis for learning analytics

Kirsty Kitto
University of Technology Sydney
Sydney, NSW, Australia
kirsty.kitto@uts.edu.au

Catherine A. Manly
City University of New York
New York, NY, United States
cmanly@gc.cuny.edu

Rebecca Ferguson
The Open University
United Kingdom
rebecca.ferguson@open.ac.uk

Oleksandra Poquet
The Technical University of Munich
Munich, Bavaria, Germany
University of South Australia
Adelaide, Australia
spoquet@gmail.com

ABSTRACT

Content analysis (CA) is a method frequently used in the learning sciences and so increasingly applied in learning analytics (LA). Despite this ubiquity, CA is a subtle method, with many complexities and decision points affecting the outcomes it generates. Although appearing to be a neutral quantitative approach, coding CA constructs requires an attention to decision making and context that aligns it with a more subjective, qualitative interpretation of data. Despite these challenges, we increasingly see the labels in CA-derived datasets used as training sets for machine learning (ML) methods in LA. However, the scarcity of widely shareable datasets means research groups usually work independently to generate labelled data, with few attempts made to compare practice and results across groups. A risk is emerging that different groups are coding constructs in different ways, leading to results that will not prove replicable. We report on two replication studies using a previously reported construct. A failure to achieve high inter-rater reliability suggests that coding of this scheme is not currently replicable across different research groups. We point to potential dangers in this result for those who would use ML to automate the detection of various educationally relevant constructs in LA.

CCS CONCEPTS

• **General and reference → Reliability**; • **Information systems → Content analysis and feature selection**.

KEYWORDS

content analysis, methodology, reproducibility, labelled data

ACM Reference Format:
Kirsty Kitto, Catherine A. Manly, Rebecca Ferguson, and Oleksandra Poquet. 2023. Towards more replicable content analysis for learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3576050.3576096

1 INTRODUCTION

How do we know that the approaches emerging from the field of learning analytics (LA) are sound? Replication is key to building support for the theories and models developed in modern research, requiring a strong consensus among members of the relevant community about how a phenomenon will manifest, across a wide range of scenarios, before a theory describing that phenomenon is trusted by the relevant research community. However, it is becoming clear that a number of results long considered statistically significant in many different fields of research cannot be replicated in new contexts [1, 19], and a dramatic restructuring of what counts as validated theory is now underway. For example, in 2015 an attempt to replicate results from psychology found that more than half the field's significant results failed reproducibility tests [4], and concerns about the dependence of artificial intelligence (AI) upon training datasets that are rarely released to the public [22] are leading to critical dialogues about the likely significance of recent results [20, 31].

One reason for these debates lies in the generation of false positive results, which can occur when a p-value that was reported as significant for one scenario does not remain significant if an experiment is replicated in a new context. Gelman and Loken [19] make a compelling argument that these failures to replicate are usually not due to malice or ineptitude, but rather to a "garden of forking paths", where the many different decisions made during a data analysis add hidden variables which are not factored into hypothesis testing, making a result seem stronger than it actually is. Alternatively, Manly et al. [30] argue that the way in which we define and then communicate the various constructs used in an analysis can cause problems. Some constructs may seem so straightforward that their operationalizations have never been specified in published quantitative research. However, this very operationalization often hides definitional differences that can lead to discrepancies in both the magnitude and direction of a construct. This set of problems has led to claims that modern research is suffering a *replication crisis* and cannot be trusted [19, 36].

These issues have led to a number of interesting initiatives to improve the replicability of results across many fields. For example,

LAK 2023, March 13–17, 2023, Arlington, TX, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576096

the Open Science movement (see e.g. https://osf.io/) encourages research groups to pre-register trials, share associated analyses, and store data publicly. Doing so helps ensure null results are reported (eliminating file-drawer problems) and, increasingly, that research data are made available for broader reuse. Several fields have long-established procedures that support reproducibility in similar ways. For example, data science research communities hold periodic competitions, such as TREC, NTCIR and MIREX[1], which have resulted in the release of a wide range of open datasets and methods, a common resource essential to many advances in the field. While Raji et al. [38] have challenged the resulting valorisation of a small number of datasets and their particular constructs, often defined in a very restrictive manner, these shareable open datasets also bring direct advantages. They help different communities to compare research results, and so better understand state-of-the-art solutions. They also support the training of new entrants to the field, who can access data, examine how it has been labelled, then develop their own sophisticated analytical methods and compare them with established baseline results.

Taken together, this body of work suggests an emerging need for the choices made in "doing LA" to be more clearly described, and for more explicit reporting about the influence of these choices on the research process and its results. This would help minimise unclear researcher degrees of freedom, which can lead to confusion, incorrect applications to other contexts, or the reporting of results as significant when they are not in fact so.

As the LAK community considers the topic of *trustworthy learning analytics* for 2023, this paper will consider some of the ways in which we might lose confidence in the results of our analyses as they are generalised beyond the group that first reported upon them. We warn of the difficulties associated with replicability that lie in wait for LA if naive implementations of content analysis (CA) are used to generate datasets that form training sets in machine-learning methods. After a brief overview of some of the approaches LA has taken to the problem of replicability (Section 2), we will consider ways in which "researcher degrees of freedom" can become embedded in research using CA (Section 2.1) before exploring some of the subtleties known to arise in reporting CA (Section 2.2). We will then consider a specific example scenario in Section 3 (that of coding for exploratory talk), and a demonstration we have used to explore the ways in which four independent "research groups" might code one dataset using CA. This will provide an understanding of how various choice points influence results generated using CA (Section 4) and of steps the LA community could take to ensure robust and replicable results. A reporting template is presented in Section 4.1, and tested in Section 4.2. Our final thoughts about this program of work and its implications for LA are discussed in Section 5.

2 REPLICABILITY IN LEARNING ANALYTICS

Attempts to reproduce existing LA-based results in new contexts have tended to fail more often than they succeed (see e.g. Farrow et al. [12], Gardner et al. [17], Hu et al. [21]). This suggests that tools based upon these methods are not currently robust enough to be used in authentic learning scenarios. While claims are often made that this is due to overfitting, a slightly different factor is also likely to be at play: the different choices made in performing a data analysis create many more degrees of freedom than are normally assumed. As a result, our significance estimates are often overly optimistic as they underestimate the degrees of freedom actually associated with an analysis.

LA has made a number of attempts to address these problems of replicability. For example, a Journal of Learning Analytics special section [10] published information about four datasets available to the broader community; the LAK data challenge [11] encouraged the sharing of more open data; and the Pittsburgh datashop[2] attempts to support more open access to educational data. More open datasets [5, 28] have gradually emerged, accompanied by a growing use of pre-registration [6–8, 25, 41, 43]; and the generation of synthetic data that can be shared [2]. Despite these ongoing attempts to encourage more open access to the data used in LA, few research groups release their data publicly, and what is released does not cover the broad range of use cases that arise in LA. This is a highly unusual situation for a field making substantial use of machine learning (ML) [18]. Around the world, groups are collecting, cleaning, exploring and analysing learning data. Each of these steps requires a range of decisions about what is counted, and indeed what counts [3]. Many decisions which influence LA procedures, coding, and results are not well documented within the group that first undertakes them, let alone for the wider LA community. It is therefore possible for different groups to make markedly different decisions when collecting, cleaning, and labelling two highly similar datasets (or potentially even the same one). Even the choices made about how an unbalanced dataset should be treated have been shown to affect the results of a ML-based classification in LA [12]. Perhaps the most sophisticated attempt to address these problems of replication arise from the MOOC replication framework (MORF)[3] which is part of an ongoing effort to produce an open-source software toolkit enabling end-to-end reproducibility [23]. Using this approach a number of replication studies have been performed, each dramatically increasing the size of the dataset used to test previous results and often demonstrating that past results fail to replicate and, indeed, contradict the original findings [17]. The MORF platform provides no direct access to underlying data, instead answering well formulated questions put to it. While this approach resolves the numerous problems with privacy that have made it impossible to share rich learner data to date, its dependence upon quantitative analysis means that this approach does not solve the problem of ensuring the trustworthiness of datasets labelled using CA.

In summary, while there is some hope for improving the replicability of LA methods that rely upon quantitative data, the qualitative labelled datasets that we use in training and testing a number of our ML methods are less well tested. While clickstream data is relatively easy to come by, more ethically fraught data can be far more difficult to source. This is particularly the case for rich textual data such as discussion forum exchanges, written evaluations of teaching, and student essays. More concerning, even if raw data of

[1] See http://trec.nist.gov/, http://research.nii.ac.jp/ntcir/index-en.html, and http://www.music-ir.org/mirex.

[2] See http://www.learnlab.org/technologies/datashop/.

[3] https://educational-technology-collective.github.io/morf/

this form can be sourced, manually labelled data resulting from a CA, whether conducted by an expert, via crowdsourcing, or by a research associate during a project is rarely publicly released. This makes it very difficult to test our results in this space for replicability. Let us now specifically consider the way in which choices made during a data analysis might impact upon the trustworthiness our results.

2.1 The forking paths of analysing educational data

How do the choices made during an analysis affect the robustness of LA results? Even before a practitioner starts an analysis, systemic factors will influence what data they can work with. National policies, the affordances of the systems used to collect data, research priorities of the lab, and the type of data they can access will all indirectly affect the analysis. Beyond these systemic influences, the way data is collected and then sampled also dramatically influences the resulting analysis. Next, the educational context in which a dataset was obtained, how subsets of data were defined, assessment details and their associated learning designs, and specifics about a student cohort all work to increase the complexity of the system being analysed. When performing the analysis, further decisions must be made, including model choice, feature selection, parameter settings, and test/train splits. The hypotheses associated with a study must be formulated, along with levels of significance and what will count as a reportable result. It is within this part of the analysis that labelling the dataset might become necessary. But what types of choice are made in labelling a dataset using CA? And how might these choices affect the robustness of the results obtained? Although the compendium of approaches to qualitative coding provided by Saldaña [39] may help content analysts develop a set of codes, reporting about labelling choices remains inconsistent. The sufficiency of standard methods followed by LA practitioners in methods reporting is unknown. We sought to understand the potential issues by asking two research questions:

Research Question 1: How do researcher degrees of freedom affect the robustness of CA results obtained by independent research groups when labelling one dataset?

Research Question 2: What steps can be taken to make CA results and methods more replicable in LA?

In order to answer these questions, we must first become more familiar with the subtleties of content analysis itself.

2.2 Content Analysis

Content analysis is a rigorous method that can be used to examine latent elements of text. It enables researchers to consider the meaning underlying manifest elements of data [37]. In the field of LA, content analysis is sometimes misinterpreted as a way of describing any analysis of content, sometimes structured, and sometimes qualitative. However, the method is more tightly defined than this. Krippendorff [27] provides a detailed description and examination of the method and its potential pitfalls in his book ‘Content Analysis: An Introduction to Its Methodology’, now in its fourth edition. He defines content analysis as “a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use” [27, p18]. In particular, Krippendorff [27, pp23-24] notes that:

(i) texts have no objective — that is, no reader-independent — qualities;
(ii) texts do not have single meanings that can be ‘found’, ‘identified’, and ‘described’ for what they are;
(iii) texts have meanings relative to particular contexts, discourses, or purposes.

These claims mean that CA requires a context in which the texts interpreted make sense and can be used to answer a set of defined questions. This context is an integral part of the analysis and is important for gaining an understanding of its results. However, it is important to be aware that this context is constructed by the researcher and is not a pre-existing objective reality that would necessarily be recognised by all contributors. Despite these issues, LA research groups applying CA to datasets in the course of performing a data analysis normally make claims about the robustness of their results via metrics such as inter-rater reliability (IRR), which is reported as if it is an objective measure, although the complexities of CA have already been pointed to by some researchers in the field [24].

Krippendorff himself discusses these issues of reproducibility, setting out requirements for any analysis that relies upon observer agreement to report upon it. These issues involve employing communicable coding patterns and training coders in their use, employing communicable criteria for the selection of coders, and ensuring coders work independently of each other. This last point is often neglected when LA researchers work to label data for ML tools and statistical tests. The coders in a research group may easily discuss the data and reach a consensus or majority decision on how it is to be coded ahead of the procedure itself. Further, interpersonal relationships and power structures in the group may shape how such negotiations and discussions impact upon these decisions, and hence the coding process. In other studies, observers work separately but consult each other when unexpected problems arise. This reinterpretation of coding instructions can mean the process loses stability over time.

In short, when establishing the reliability and validity of CA, it is important to provide a detailed account of the methods employed. De Wever et al. [9] provide a strong case for this in their analysis of 15 CA schemes applied to online asynchronous discussion. They found a number of problems that made it impossible to judge the reliability of the results in most cases given the way the schemes had been applied, concluding that:

> “systematic coherence between theory and analysis categories, a grounded choice for the unit of analysis, and information about the (inter-rater) reliability and procedure are necessary conditions for applying content analysis in the context of a sound research methodology” De Wever et al. [9, p6].

More specifically, they identified information about the method that should be included when writing up results: (i) size of, and method used to create a reliability sample; (ii) justification for using that method; (iii) relationship of sample to full sample; (iv) number of coders and whether they included the researchers; (v) amount of coding conducted by each coder; (vi) approximate amount of training to reach the stated reliability levels; (vii) where and how people can learn more about the instrument; (viii) inter-rater reliability level for each variable. It is rare to see all these elements included in a LA publication that makes use of CA (although at least some of them are *usually* included).

IRR is the common method used to demonstrate that a CA is robust enough to be considered reportable and so reliable for ML-based studies. However, there is no one commonly agreed measure for IRR, making the objectivity often attributed to this measure less straightforward than it might appear [9]. Different research groups use and report different IRR measures. A basic measure is percentage agreement, which compares total number of agreed codes with total codes. This measure can be severely skewed if coders agree on a large number of null codes, only disagreeing on the few cases that contain the relevant construct. Scott's pi and Cohen's kappa compare the labelling between two coders; both take into account the possibility of chance agreement. Fleiss's kappa compares the work of more than two coders, expressing the extent to which the observed agreement exceeds what would be expected if all raters had coded at random. Krippendorff's alpha is more complex relative to other metrics but has the advantages that it works for any amount of data, any number of coders, and ignores missing data. Reporting on IRR is further complicated if there is no agreed unit of analysis, so coded text overlaps but is not identical [40]. With so many options possible, it often is difficult to gauge just what it is that a team of coders are agreeing on, and where their disagreements are located, from a reported IRR metric. However, moving beyond a simple report of the metric to explicitly reveal these divergences can highlight where further conceptual clarity is necessary. This can help to prevent automation of the false positives or negatives during ML. A lack of transparency about the CA process and the effect this may have on the reported quality metrics raises concerns about the trustworthiness of text labelling efforts and their further application to ML. We shall now make use of a case study to explore these issues in more depth.

3 CASE STUDY: CODING FOR EXPLORATORY TALK

LA has made use of datasets coded using CA to classify a number of constructs, including cognitive presence [26], helpseeking behaviour [5] and exploratory dialogue [14]. We had access to the exploratory talk dataset described in Ferguson and Buckingham Shum [14], which has previously been analysed by Ferguson et al. [15] using ML, suggesting that the detection of exploratory talk can be automated, and making it a prime candidate for a LA replication study.

3.1 What is exploratory talk?

Mercer and his colleagues [32–35] distinguished three social modes of thinking used by groups of learners: disputational, cumulative, and exploratory. They claimed exploratory dialogue is that which instructors consider most educationally desirable [42]. It can be found in both online and offline learning environments [13, 16], providing an indication that learning is taking place and that learners are going beyond a simple accumulation of ideas. Mercer and Littleton [34] described the appearance of exploratory talk in a school environment as follows:

> "Exploratory talk represents a joint, coordinated form of co-reasoning in language, with speakers sharing knowledge, challenging ideas, evaluating evidence and considering options in a reasoned and equitable way. The children present their ideas as clearly and as explicitly as necessary for them to become shared and jointly analysed and evaluated. Possible explanations are compared and joint decisions reached. By incorporating both constructive conflict and the open sharing of ideas, exploratory talk constitutes the more visible pursuit of rational consensus through conversation." [34, p62]

In contrast to the work above, which collected data about exploratory talk in a face-to-face environment, Ferguson et al. [15] used data collected from a two-day online teaching and learning conference organised by The Open University in 2010. This was gathered using the web-conferencing tool Elluminate and included all synchronous text-based discussion among the participants. The majority of participants were higher education researchers and practitioners from around the world. Asynchronous discussions in the data took place in relation to recorded presentations that were not captured in the dataset. Ferguson et al. [15] considered each individual post to the chat to be a turn in the dialogue coded according to four sub-categories:

Challenge: identifies that something may be wrong and in need of correction. It may involve calling into question, account, dispute, finding fault with, raising an objection, putting forward an opposing view or proposing revision.

Evaluation: has a descriptive quality. It may involve assessment, expressing a concept in terms of something already known, appraisal or judgement.

Extension: builds on, or provides resources that support, discussion. It may involve increasing the range of an idea or concept, applying it to a new area, taking the same line of argument further, proposing, developing or linking to related resources.

Reasoning: includes thinking an idea through. It may involve discussion, argument, explanation, inference, asking questions about content, reaching a conclusion, working ideas out in a logical manner, changing your view in the light of arguments presented, justifying your position, requesting additional resources to support understanding.

Postings identified as containing one or more of these sub-categories were taken to be instances of exploratory dialogue.

3.2 Experimental Demonstration 1: CA across independent research groups

To explore what issues may arise in using CA to generate a labelled dataset that can be used in ML and LA, we conducted an experimental demonstration (which we will call an experiment here for conciseness). This involved the authors of this paper taking the role of four independent researchers, so representing members of four different "research groups" who were working to code the data from one session of the Elluminate conference. This sample of "research groups" embodies different subject matter expertise, as well as differing knowledge of LA research community norms and methods, and a variety of background knowledge concerning CA and ML. Importantly, none of the coders were novices in applying CA. One participant was a non-native speaker of English, but the other three were native English speakers. We sought to investigate Research Question 1 by being explicit about the choices made throughout our analysis, working to determine how much divergence might result if no communication was used to harmonise the choices made by independent researchers while performing CA (as is likely to become the case as more groups attempt to replicate or build upon previous LA results). In short, Experiment 1 was designed to illustrate what can happen if research teams acting independently of one another attempt to code the same construct using the typical amount of knowledge provided via public coding instructions and published literature. This scenario reflects a situation where future coders rely on prior published studies without having inside information about how that coding was conducted. Coding without consultation also addresses Krippendorff's assertion that coding should occur independently in order to ensure stability, assuming that the instructions to coders provide enough information to produce stable codes.

This experiment was performed in October 2019. Our four independent researchers each received: a coding scheme for exploratory talk, a file describing the selection and description of the dataset, a code descriptions file, and a de-identified spreadsheet of data to be coded. The researchers exchanged several articles on the underlying theory, and separately engaged with literature discussing exploratory talk. They had a further asynchronous exchange confirming the unit of analysis to be coded — defining this as a single posting made by one individual. In order to maintain independence, the coders did not communicate the overarching procedure they were planning to use, and did not have any further discussions clarifying any of the codes as they worked to label the dataset.

What further choices were made by our independent "research groups" once they began to study the dataset used in this experiment? Table 1 explicitly lists the decisions made by each coder as they operationalised the coding scheme of exploratory talk and then worked to label the dataset provided. Note that these extra details were recorded privately, and not shared between the groups until after the coding had been completed.

Codes from each of the four "research groups" were then analysed using four inter-rater reliability (IRR) measures. Table 2 reports the percent agreement, Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha (all discussed in section 2.2). In the case of each sub-category (i.e., challenge, evaluation, extension, and reasoning) the percent agreement is much higher (almost perfect), while the kappa and alpha IRR coefficients are much lower, generally scoring a fair rating. This discrepancy arises because the data coded have a high number of off-topic posts that do not constitute exploratory talk. High agreement on these uninteresting items skews the percent agreement, while not fully reflecting the actual agreement on the main exploratory talk construct. Notably, the kappa and alpha values are more in line with expectations, as they provide a more accurate measure of agreement between coders in the case where one large category dominates. In many cases, the IRR value of the subcategories is very close to the 0.2 cutoff for being considered slight instead of fair, reflecting substantially less agreement between the coders for the sub-categories than arises for the higher level exploratory talk construct itself (which scores above 0.5).

Drilling further into the data generated in this experiment provides some important insights. Firstly, some very different results were obtained about how many posts should be designated as exploratory (see Table 3). Note that two of our coders (CM and RF) have adopted a particularly generous stance in what they classify as exploratory talk, whereas the other two (KK and SP) are far more conservative. This is interesting given that Table 1 shows KK explicitly claiming to have adopted a generous stance, highlighting that the self-perceptions of a group in reporting their results may differ substantially from the reality. In short, while IRR is generally thought to indicate coder agreement, it very much depends upon which metric is reported. If a simple measure for overall agreement is recorded using a single statistic then this can mask substantial variation in both method and sub-category coding, especially *between* groups.

4 DISCUSSION: CHOICE POINTS AND THEIR IMPACT

CA is often driven by example text snippets, which are carefully examined by coders as they attempt to label a dataset. However, as we explored the results of Experiment 1 it became apparent that we had made a number of implicit choices beyond those that were publicly available in the scheme. This is despite the fact that we had followed the best practice approach advocated by De Wever et al. [9] (discussed in Section 2.2). These extra *choice points* impacted upon a number of key parameters that, in turn, affected our labelling process. Note that reporting IRR alone (especially at the top category of exploratory talk) would do very little to report on these extra choice points, and so would not provide a true reflection of how potentially untrustworthy the results are. This problem of vague descriptions of the CA process is not new, even in the field of education. As early as 2006, Strijbos et al. [40] explored the ways in which different choices made about the unit of analysis could impact upon the reliability of coding over collaborative learning data. This led the authors to propose an alternative, more rigorous, method for defining a unit of analysis, but we are unaware of any research group in the LA community making use of this procedure. Being aware of this challenge, we took care in defining the unit of analysis (an individual post) resulting in no discrepancies for this choice point in Table 1. However, a number of other decisions were made implicitly. Our second experiment sought to remedy this problem and to create a process where these decisions were made explicit.

4.1 A proposed solution: reporting requirements for reproducibility in CA

It appears that more robust reporting is required to improve IRR between our independent "research groups". After the results of Experiment 1 we carefully explored our processes in an attempt to understand our poor IRR values. Combining results from Experiment 1 with further analysis of the literature, we synthesized our

Table 1: The coding process followed by each "research group", noting significant differences which are rarely reported upon.

	Coder 1 (KK)	Coder 2 (CM)	Coder 3 (RF)	Coder 4 (SP)
What was your process?	Explored first 100 entries to get a feel for data. Worked sequentially through each entry, coding it in one pass. Some sentences were identified as miscoded after reading ahead — line was recoded (note left). No repeat check.	Skimmed through data to determine character of entries. Worked sequentially through each line of text in one pass. Decided on approach to coding and tagged everything during this pass. No repeat check.	Had already coded data in previous studies and was familiar with instructions. Worked sequentially through each line of text in one pass. Added note where disagreement was anticipated. Some sentences were identified as miscoded after checking which talk was being discussed — line was recoded (note left). No repeat check.	Read the entire piece for coding and colour-coded elements that appeared to constitute connected units (differentiating text to be coded and not). Refined scheme with additional keywords based on impressions from first read. Coded colour-coded text with as little deliberation as possible. Spot checked some of the codes a week later.
What order did you code in?	Coded line by line. Determined whether line satisfied at least one subcategory (more than one allowed). If at least one subcategory was present then line was deemed exploratory. Not all posts about subject of conference were denoted exploratory.	Coded line by line. Determined whether line was exploratory. If exploratory, line was coded according to the appropriate subcategory. All posts about the subject of conference were given subcategories and called exploratory talk.	Noted where I thought that talk was cumulative (another form of dialogue) rather than exploratory. After coding more than half checked the context — what speakers were talking about at specific times. This prompted some recoding as I realised some comments I had previously considered off topic were actually exploratory.	First parsed the colour-coded text into sub-conversations. Interpreted exploratory talk in relation to the context provided across these larger contextual pieces (often comprised of several separate units of text by different people). Did not consider off-topic units that were not colour coded as part of the context for interpretation.
How did you define exploratory talk?	Tried to find a category the post satisfied. If one could be found then the post was labelled exploratory.	Determined whether talk was part of a conversation about the material being discussed in the session. (If it was not, the post was considered non-exploratory.)	Used the definition of exploratory in the instructions for coders. Identified whether a contribution was exploratory according to that definition, then decided which subcategories were appropriate.	Talk that included clear and explicit reference to the scheme within the colour-coding. If deliberating over whether the text *could* be exploratory occurred it was coded as non-exploratory (as lacking explicitness and so hard to replicate.)
What did you interpret exploratory talk against?	Had difficulty due to the topic of the conference, which could be seen as including environments in which learning occurs. Adopted a generous stance - if a subcategory could be found to code something not just about using Elluminate then it was included.	Talk engaged in by the participants where there were either: (a) Direct exchanges between participants in their posts, or (b) Talk in discussion about something the unseen speaker said.	Excluded exploratory dialogue that related to social issues (the other participants and the group as a whole) and about the tools and processes available to participants (for example, how to use Elluminate effectively).	Against a set of other separate text pieces that comprised an obvious conversation about the subject. Anything where it was possible to see relevant subject dialogue unfold around it was treated as ET.
How did you assign subcategories?	Read a comment and then considered it against each of the four subcategories. Some comments were coded more than once.	First determined whether there was an obvious category. If not, then looked through the category descriptions again and thought about which matched the best. If no obvious category, then typically used Extension.	Read a comment. If it matched the definition of exploratory dialogue, I considered it in terms of each of the four sub-categories. Some comments were coded more than once.	Distinguished the categories prior to coding, to ensure a clear understanding of where they differed. Tried not to consult the scheme when coding, only if in doubt.

divergent choice points into a preliminary template, presented in Table 4. This table lists eight choice points that we considered critical, along with a set of decisions that a research group might make in responding to them. While many of these decisions may appear self-evident once stated, some are more subtle when it comes to the CA process. We elaborate here on some of the nuances which we considered during the process of creating this table.

Firstly, Choice point 1 introduces some units of analysis that are less common than is the norm in LA. The options encourage explicit descriptions of the data. The different choices for the unit of analysis also demonstrate the variety and complexity of the interactions in forums or chats, with either text, emojis, audio, or video. Our guidelines targeted primarily text-based sources such as posts and related transcripts since these are common in CA conducted by the LA community. Choice point 2 is about whether multiple codes can be assigned to the unit of analysis. Choice point 3 refers to the process of coding itself. Both can significantly impact IRR in CA depending on how much data directly relates to the construct of interest (i.e. how much noise exists), or how long the unit of analysis is (which may open the possibility that multiple codes might be observed). Such choices may have sizable impact on IRR when coding

Table 2: Inter-rater reliability for Experiment 1 using a set of different measures. Note: IRR calculations for subgroups made with Stata's kappaetc command. Benchmark interval scale from Landis and Koch [29] . All results p<0.000.

	Challenge		Evaluation		Extension		Reasoning		Exploratory Talk	
IRR Measure	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval
Percent Agreement	0.931 (0.007)	Almost Perfect	0.892 (0.009)	Almost Perfect	0.832 (0.010)	Almost Perfect	0.899 (0.003)	Almost Perfect	0.818 (0.010)	Almost Perfect
Cohen/Conger's kappa	0.205 (0.034)	Fair	0.210 (0.037)	Fair	0.333 (0.029)	Fair	0.219 (0.037)	Fair	0.558 (0.024)	Moderate
Scott/Fleiss' kappa	0.201 (0.034)	Fair	0.207 (0.037)	Fair	0.322 (0.031)	Fair	0.216 (0.037)	Fair	0.553 (0.025)	Moderate
Krippendorff's alpha	0.201 (0.034)	Fair	0.207 (0.037)	Fair	0.322 (0.031)	Fair	0.216 (0.037)	Fair	0.553 (0.025)	Moderate
Observations	667		667		667		667		667	

Table 3: Number of posts classified as exploratory (Yes) or not (No) by our four independent coders in Experiment 1.

	KK	CM	RF	SP
Yes	130	235	260	133
No	537	432	407	534
%	19.5	38.9	39.0	19.9

datasets from vastly different sources. For example, the difference between using a null code (so that every unit of analysis gets a label) and allowing no code at all if a unit of analysis fails to match a sub-code might substantially change how much data is considered relevant for calculating IRR, and hence its resulting value. Choice point 4 asks if some parts of the data have been excluded, which can also lead to some subtleties in reporting. In our case, the dataset had previously been cleaned of emoticons, which meant that a number of units of analysis were empty as they contained only these symbols. Since the dataset we received was already cleaned, technically, not all the text was coded by us. However, the dataset shared with our teams was this cleaned dataset presented as whole. In other words, we coded all of it, although what we received was not all the data in their entirety. We decided that we should report Choice 4 as (i). This type of decision making is rarely reported upon, but can have a significant impact upon results if a different group were to decide otherwise. Choice point 5 resulted in a number of discrepancies in Experiment 1 (see Table 1); some "research groups" coded as they went sequentially and others read the entire dataset before commencing coding. Choice point 6 refers to how each coder treated the context of the interactions. This appears to have created substantial difficulties for our analysis. Upon consideration, the questions "How did you define exploratory talk?" and "What did you interpret exploratory talk against?" that appear in Table 1 are both attempting to understand how the different "research groups" were working to understand the context of a post. In Experiment 1, the inferred context was treated in a very different manner by each of the four "research groups", which is likely to be a key source of divergence between our results. While implied context in the documents describing the dataset was included, different coders treated it differently. Similarly, Choice point 7 that addresses decisions for coding units of analysis that are difficult to interpret caused significant discrepancies in Experiment 1. We believe this is likely due to divergences in how the four different coders interpreted ambiguous posts (i.e. posts that did not directly fit into a unit of analysis.) For example, a participant posted, "Sorry - please see my contribution about this experience in acliud called E-learning Africa" and then followed up in a separate chat statement saying "I mean 'cloud'." All coders agreed the first post constituted extension, but half coded the second as null. This raises a related point that we did not consider in our analysis: At what point is the scheme used for coding fixed and not open to change? Performing no repeat coding of the scheme represented an assumption that exploratory talk is a well-defined construct, which left us with little provision for refining it when necessary.

Beyond the choices in Table 4, a well performed CA should also report the points of disagreement that arose between coders, as these can illuminate key uncertainties in the construct being labelled. For example, if a majority of disagreements revolved around one subcategory then it would be important to highlight this problem as it points to an ambiguity in the scheme. Similarly, the background of the coders themselves (e.g. their native language and proficiency in CA) may contribute to the areas of disagreement, but the LA community has no standards for reporting these details.

4.2 Experimental Demonstration 2: Aligning between groups using detailed methodological writeup

Experiment 2 was performed in September 2022. Having attempted to formalise the choice points identified as critical in Experiment 1, we agreed on a set of choices we would adhere to as we recoded the original dataset. The choices made are underlined in Table 4. A three-year gap between coding episodes made us confident that there would be little recall of original practices, and care was taken not to inspect our previous data or results.

The IRR across the four coders for Experiment 2 is reported in Table 5. Despite a reduction in percentage agreement, most other IRR statistics improved. The exception to this was the Challenge subcategory, which now had considerably more disagreement among our independent "research groups". Table 6 drills into these results in more detail, showing where the various pairs of coders improved in their agreement over the various sub-categories. It is difficult

Table 4: Making choices in performing a content analysis over discussion datasets explicit. If a decision is made that is not identified in this table, its parameters and decision-making process should also be specified. Decisions made in Experiment 2 are underlined.

Choice point	Reportable decisions made in preparing for analysis.
1: What unit of analysis is coded?	(i) Complete utterance (eg a forum post).
	(ii) Sentence within an individual's utterance at one point in time.
	(iii) Meaning unit within a person's utterance at one point in time. This could be less than a sentence or run over multiple sentences.
	(iv) Meaning unit within several utterances made by the same individual at multiple time points, one after the other.
	(v) Meaning unit within several utterances made by the same individual at multiple time points, which may be interleaved with another individual's discourse.
	(vi) Meaning unit across several utterances of different individuals who contributed multiple times, potentially interleaved with other meaning unit(s) of one or several other individuals.
2: Can a unit of analysis have multiple codes? (Answer all sub-queries)	(i) If an item cannot be coded does it receive a null code? (YES/NO).
	(ii) If an item can be coded in multiple ways does it receive a single mixed code? (YES/NO).
	(iii) What is the maximum number of codes that a unit of analysis can receive? Justify your decision. (There are 4 codes maximum. Exploratory talk has four subcodes which can all be applied, a null receives one code.)
3: How comprehensive is the coding?	(i) Every unit of analysis is coded (some may receive a 'Null' or 'Mixed' code).
	(ii) Only units of analysis that relate to the construct of interest receive a code.
4: Are certain contributions excluded *a priori*?	(i) The complete dataset is coded.
	(ii) Text produced in certain ways, such as by bots or a computer program, and/or certain elements (e.g., emoticons, unusual punctuation, URLs, GIFs, symbols, formulae) are excluded from the dataset. Specify which elements are excluded from the dataset and explain the decision
	(iii) Text produced in certain ways, such as by bots or a computer program, and/or certain elements (e.g., emoticons, unusual punctuation, URLs, GIFs, symbols, formulae) are excluded from analysis but included within the dataset for context. Specify which elements are excluded from analysis and explain the decision.
5: How much text is reviewed before coding?	(i) No pre-reading. Read and code one unit of analysis at a time.
	(ii) Read all units of analysis within a defined set (e.g. a thread) that is larger than the unit of analysis, then code within that group/thread before moving to the next.
	(iii) Read all units of analysis, then code one at a time.
	(iv) A repeat coding strategy is adopted. Describe your process.
6: Is context taken into account when coding?	(i) Only the contents of the unit of analysis are coded.
	(ii) Context referred to and/or implied (state which) by the individual is taken into account.
	(iii) Context referred to and/or implied (state which) within a specified section of the dataset, such as a thread, is taken into account.
	(iv) Context referred to and/or implied (state which) within the entire dataset is taken into account.
	(v) Contextual information not included within the dataset but available to one or all coders (state which) is taken into account.
7: How are decisions made about units that are difficult to interpret? (Answer all sub-queries)	(i) If a code is not a direct fit for a unit of analysis, it is not applied and coding proceeds according to Choice 2. YES/NO
	(ii) If a unit of analysis includes an element related to a code or the opposite of that code (e.g., the code is 'dark' and the text refers to 'bright light'), it receives that code. YES/NO
	(iii) Units of analysis can be left uncoded and flagged for group discussion. YES/NO
8: Do coders repeat their coding?	(i) Yes. Identify the required number of repetitions.
	(ii) No. Code all cases as they are read, without repeating the process.

to see any strong patterns of improvement resulting from the application of our more formalised coding procedure at this level of analysis. In the right circumstances it would be possible to report on the agreement between some of these groups as significant. For example, the improvement in agreement between coders KK and RF is notable. However, the failure to achieve consistent agreement across the other coders in Table 6 suggests this cannot be considered as anything other than a false positive. Far more evidence would be required to report upon an improvement attributable to our proposed requirements in Table 4. Importantly, note that the IRR of individual coders *with themselves* is poor. Table 7 shows a substantial range of disagreement between each of the coders over Experiments 1 and 2, with subcategories never scoring above a Moderate agreement level, although the identification of exploratory talk as a general category appears to have been somewhat more consistent. While this could be due to the application of new more specific coding instructions via Table 4, it may also be that the scheme is not well enough defined for consistent application.

It is important to realise that this result has only emerged from our attempt to replicate previously reported coding, which had been assumed as well specified and hence stable over time.

Two decisions that appear to affect the IRR achieved in Experiment 2 the most are Choice Points 6 and 7 (see Table 4). As an example for Choice 6, three coders coded the post "@Jose or Escher!" as null, but one chose evaluation. Given the conversational context (following the posts "@Jose How do you QA a passion?" and "@Bill - ask the learners!" which all four coders agreed were both exploratory), the participant was seen to be making an evaluative comment about complexity based on "MC Escher's circular drawings" which was clarified in a later post also seen as exploratory only by the same coder. While we anticipated that making Choice 6 more explicit would help, we consider it likely that more work is required to clearly delimit how context is to be interpreted as this choice point still appears to be causing problems in Experiment 2. These data also point to the complexity of the task itself — without understanding the cultural reference to Escher it is very hard to interpret the comment with the context as evaluative. Similarly, our four independent coders appear to have interpreted Choice 7 quite differently, resulting in divergences (especially for Choice 7(i)). We consider these choice points to be a key weakness in attempting to generate reproducible CA results across different research groups, and our study has yet to make them explicit enough to control for coder divergences. We believe both choice points likely require significant further investigation for the field of education. Finally, the choice points we defined in Table 4 did not clarify whether to begin coding by identifying the main construct or each of the four subcategories. However, given this decision's influence, distinguishing between starting from higher level or lower level coding perhaps should be added to our requirements. Experiment 2 provided further support that identifying subcategories where significant disagreement occurs, such as challenge, might signal subareas of the overall construct that might be fruitfully reconsidered over time. Indeed, Experiment 2 has led us to seriously reconsider whether the subcategories of exploratory talk make sense at all. It may make more sense just to define the one overarching category (i.e. exploratory talk).

5 CONCLUSIONS

Perhaps the most obvious conclusion to draw from our two experiments is that we have failed to reliably replicate the exploratory talk coding scheme used to generate the results presented by Ferguson et al. [15]. This failure occurs at the level of the initial coding of the dataset. As a result we do not believe that it is currently possible to code exploratory talk for this dataset in a manner that is consistent enough to support strong IRR across independent research groups. Thus, at least one study in LA that aims to automate the detection of an educational construct is suspect.

Returning to our original Research Questions, what have we learned in this study? Research Question 1 asked what types of researcher degrees of freedom creep into a CA, and how this might affect the results obtained by independent research groups. To answer this question we used an authentic experiment to simulate a process by which independent research groups might use CA to code a dataset that they hope to use in ML. We saw that a wide array of choice points emerged during this process, which both diminishes IRR scores and, more importantly, would lead to different features in a dataset being emphasised if a move were then made to automate detection of various codes. Note also that the same coders could exhibit quite different results over time. The LA community should work hard to ensure that our methods are not becoming overly tuned via the various choices that different research groups are making around the world.

Research Question 2 sought to identify steps we might take as a community to ensure that our results are replicable across different research teams. The scheme in Table 4 was trialled as a mechanism by which LA researchers might be able to explicitly record some of the important choice points that are usually treated implicitly. Widespread use of a standard CA reporting template like this would lead to more relevant information being communicated by research groups using CA. This would help to ensure that LA results are more likely to be reproducible. It would also help to ensure that our ML models are robust as they would be based upon more trustworthy data. However, attempting to use Table 4 in Experiment 2 yielded mixed results. While some general improvement can be identified, there are still too many discrepancies emerging between our coders for this to be considered a robust tool. Further research and development is needed. Such work may include investigations of the areas of disagreements and potential refinement of constructs, or the creation of more refined reporting tools that can help reduce researchers degrees of freedom.

It may be possible to claim that the problems identified in this paper result from a difficult CA scheme that is applied to a dataset collected in a learning scenario that was not designed to elicit exploratory talk. We acknowledge that other constructs and datasets *might* achieve more robust results between independent research groups. However, to the best of our knowledge this is not something that the LA community has tested. The sheer complexity of educational constructs suggests that problems are very likely to occur if our methodology were repeated in other scenarios. We encourage more research groups to undertake this test. Other readers might question our lack of conversations to achieve convergence. While the standard way of achieving convergence within an existing research group relies upon conversations between coders,

Table 5: Inter-rater reliability for Experiment 2 using different measures. IRR improvements are noted in bold. Note: IRR calculations for subgroups made with Stata's kappaetc command. Benchmark interval scale from Landis and Koch [29] . All results p<0.000.

	Challenge		Evaluation		Extension		Reasoning		Exploratory Talk	
IRR Measure	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval	IRR (SE)	Bench. Interval
Percent Agreement	0.928 (0.007)	Almost Perfect	0.860 (0.009)	Almost Perfect	0.776 (0.010)	Substant.	0.859 (0.010)	Almost Perfect	0.795 (0.010)	Substant.
Cohen/Conger's kappa	0.194 (0.037)	Slight	**0.289** (0.032)	Fair	**0.347** (0.027)	Fair	**0.391** (0.031)	Fair	**0.586** (0.021)	Moderate
Scott/Fleiss' kappa	0.192 (0.038)	Slight	**0.289** (0.037)	Fair	**0.342** (0.027)	Fair	**0.391** (0.031)	Fair	**0.584** (0.022)	Moderate
Krippendorff's alpha	0.192 (0.038)	Slight	**0.289** (0.037)	Fair	**0.342** (0.027)	Fair	**0.391** (0.031)	Fair	**0.584** (0.022)	Moderate
Observations	667		667		667		667		667	

Table 6: Pairwise agreement (Cohen's kappa) between coders for the two experiments (Exp1 and Exp2) across Exploratory Talk and its four subcategories. If the agreement improved in Experiment 2 then the figure appears in bold. All results p<0.000.

	KK-CM		KK-RF		KK-SP		CM-RF		CM-SP		RF-SP	
	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2
Challenge	0.349	0.180	0.174	**0.263**	0.095	**0.272**	0.165	0.084	0.227	0.129	0.048	**0.137**
Evaluation	0.293	**0.387**	0.205	**0.314**	0.175	**0.312**	0.147	**0.345**	0.260	0.177	0.170	**0.196**
Extension	0.393	0.359	0.252	**0.463**	0.391	0.283	0.410	**0.477**	0.296	0.265	0.288	0.285
Reasoning	0.209	**0.344**	0.203	**0.594**	0.273	**0.394**	0.229	**0.354**	0.198	**0.267**	0.215	**0.352**
Exploratory Talk	0.579	0.571	0.494	**0.620**	0.692	0.654	0.593	**0.711**	0.548	0.468	0.478	**0.506**

Table 7: Pairwise agreement (Cohen's kappa) of the 4 coders with *themselves* for the two experiments (Exp1 and Exp2) across Exploratory Talk and the four subcategories that were coded. All but two results p<0.000, with the p>0.000 results marked (†).

	KK	CM	RF	SP
Challenge	0.326	0.222	0.084†	0.258
Evaluation	0.315	0.488	0.245	0.098†
Extension	0.344	0.574	0.500	0.185
Reasoning	0.224	0.322	0.139	0.231
Exploratory Talk	0.423	0.699	0.616	0.327

this is not a scalable long-term solution. People move on, data are lost, and new entrants appear in a field. As the experiments here show, different research groups are quite likely to vary in their use of a detailed and carefully developed scheme if these conversations between coders go unreported. We must start to develop more rigorous protocols that make it possible to cross check codings and the methods used to obtain them. While conversations between coders will no doubt remain important within research groups and their immediate collaborators, being explicit about key methodological choice points that affect reproducibility should become standard practice in our field. Articulating a map of choice points should become a fundamental element of any LA publication involving CA (and quite likely for all analyses that we undertake). Education is a complex field, which means we must work to make our implicit biases and assumptions explicit, especially when they serve as input to ML applications. Not to do so leaves us open to the problems of replication that are besetting other fields.

We note that the theoretical origins of a carefully developed construct such as exploratory talk embrace a qualitative orientation that invites questioning the validity of appropriating such a construct for CA in a quantitative context. Indeed Mercer, who led the work that identified cumulative, disputational, and exploratory talk, wrote:

> "We have had no wish to reduce the data of conversation to a categorical tally, because such a move into abstracted data could not maintain the crucial involvement with the contextualised, dynamic nature of talk which is at the heart of our sociocultural discourse analysis. Rather, the typology offers a useful frame of reference for making sense of the variety of talk" [33, p146].

But this categorical tallying is precisely what happens when data are labelled for ML. Such constructs developed for a qualitative context may not work as validly within an "equivalent" quantitatively oriented environment where context is not at the fore. This points to a qualitative/quantitative tension which is underexplored but likely to become a significant issue for LA as it attempts to apply quantitative methods to concepts developed using qualitative approaches. Being explicit about choice points relevant to increasing implementation clarity and reproducibility, as Table 4 encourages, seems likely to move LA closer to replicable construct analysis and therefore more widely meaningful practice. However, we are yet to find a strong reporting framework.

As with all data analyses, the process of conducting content analysis is complex, involving many choices and decision points. The complexity of the texts that educational researchers might want to analyse using this method can make the results that they obtain highly contextual. This subtlety is often lost when LA researchers make use of CA to generate labelled datasets for use in ML. This paper has worked to make the various choice points inherent in using CA to label a dataset explicit, highlighting places where divergences between independent research groups are likely to impact upon the replicability of the coding scheme that is generated. Our exploration of the challenges associated with reproducing LA results highlights an often recognised, but rarely acted upon, need for open data in LA. As well as being less prone to issues of replicability, communities that have access to open data tend to develop faster, as a lower barrier exists for new entrants. In a similar way, companies and research groups with access to large closed datasets tend to excel. The LA community must consider which path we would like to follow — towards open and widely accessible data for all? Or continuing the more restrictive model that has dominated current practice? In this paper we have presented evidence that a closed approach can lead to poor outcomes that are not reproducible in the complex world of education in the wild. Far more work remains to be completed before we can feel confident that CA results are robust enough to be reliably utilised in ML-based approaches that apply across broad contexts. We hope that this paper provides one small step towards trustworthy LA that stands on the firm foundation of highly replicable results.

REFERENCES

[1] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452.

[2] Alan Mark Berg, Stefan T Mol, Gábor Kismihók, and Niall Sclater. 2016. The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics* 3, 1 (2016), 107–128.

[3] Simon Buckingham Shum and Rose Lukin. 2019. Learning Analytics and AI: Politics, Pedagogy and Practices. *British Journal of Educational Technology* 50, 6 (2019), 2785–2793.

[4] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[5] 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.

[6] Dan Davis, Ioana Jivet, René F Kizilcec, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2017. Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference.* ACM, 454–463.

[7] Dan Davis, René F Kizilcec, Claudia Hauff, and Geert-Jan Houben. 2018. The half-life of MOOC knowledge: a randomized trial evaluating knowledge retention and retrieval practice in MOOCs. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge.* ACM, 1–10.

[8] Dan Davis, René F Kizilcec, Claudia Hauff, and Geert-Jan Houben. 2018. Scaling Effective Learning Strategies: Retrieval Practice and Long-Term Knowledge Retention in MOOCs. *Journal of Learning Analytics* 5, 3 (2018), 21–41.

[9] Bram De Wever, Tammy Schellens, Martin Valcke, and Hilde Van Keer. 2006. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & education* 46, 1 (2006), 6–28.

[10] Stefan Dietze, George Siemens, Davide Taibi, and Hendrik Drachsler. 2016. Editorial: Datasets for Learning Analytics. *Journal of Learning Analytics* 3, 2 (2016), 307–311.

[11] Hendrik Drachsler, Stefan Dietze, Eelco Herder, Mathieu d'Aquin, and Davide Taibi. 2014. The learning analytics & knowledge (LAK) data challenge 2014. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge.* ACM, 289–290.

[12] Elaine Farrow, Johanna Moore, and Dragan Gašević. 2019. Analysing discussion forum data: a replication study avoiding data contamination. In *Proceedings of the 9th international conference on learning analytics & knowledge.* 170–179.

[13] Rebecca Ferguson. 2009. *The construction of shared knowledge through asynchronous dialogue.* Ph. D. Dissertation. The Open University.

[14] Rebecca Ferguson and Simon Buckingham Shum. 2011. Learning analytics to identify exploratory dialogue within synchronous text chat. In *Proceedings of the First International Conference on Learning Analytics and Knowledge.* ACM, 99–103.

[15] Rebecca Ferguson, Zhongyu Wei, Yulan He, and Simon Buckingham Shum. 2013. An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge.* ACM, 85–93.

[16] Rebecca Ferguson, Denise Whitelock, and Karen Littleton. 2010. Improvable objects and attached dialogue: new literacy practices employed by learners to build knowledge together in asynchronous settings. *Digital Culture & Education* 2, 1 (2010), 103–123.

[17] Josh Gardner, Christopher Brooks, Juan Miguel Andres, and Ryan Baker. 2018. Replicating MOOC predictive models at scale. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale.* 1–10.

[18] Dragan Gašević, Carolyn Rose, George Siemens, Annika Wolff, and Zdenek Zdrahal. 2014. Learning analytics and machine learning. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge.* ACM, 287–288.

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

[20] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S Greene, et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E14–E16.

[21] Yuanyuan Hu, Claire Donald, and Nasser Giacaman. 2022. A revised application of cognitive presence automatic classifiers for MOOCs: a new set of indicators revealed? *International Journal of Educational Technology in Higher Education* 19, 1 (2022), 1–21.

[22] Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis. *Science* 359, 6377 (2018), 725–726.

[23] Stephen Hutt, Ryan S Baker, Michael Mogessie Ashenafi, Juan Miguel Andres-Bray, and Christopher Brooks. 2022. Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology* (2022). https://doi.org/10.1111/bjet.13231

[24] Simon Knight and Karen Littleton. 2015. Discourse-centric learning analytics: mapping the terrain. *Journal of Learning Analytics* 2, 1 (2015), 185–209.

[25] Kenneth R Koedinger, Ryan SJd Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. 2010. A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining* 43 (2010), 43–56.

[26] 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 and Knowledge.* ACM, 15–24.

[27] Klaus Krippendorff. 2004. *The content analysis reader* (2 ed.). Sage.

[28] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open university learning analytics dataset. *Scientific data* 4 (2017), 170171.

[29] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[30] Catherine A Manly, Ryan S Wells, and Suzan Kommers. 2018. The influence of STEM definitions for research on women's college attainment. *International journal of STEM education* 5, 1 (2018), 45.

[31] Scott Mayer McKinney, Alan Karthikesalingam, Daniel Tse, Christopher J Kelly, Yun Liu, Greg S Corrado, and Shravya Shetty. 2020. Reply to: Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E17–E18.

[32] Neil Mercer. 2002. *Words and minds: How we use language to think together.* Routledge.

[33] Neil Mercer. 2007. Sociocultural discourse analysis: Analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics and Professional Practice* 1, 2 (2007), 137–168.

[34] Neil Mercer and Karen Littleton. 2007. *Dialogue and the development of children's thinking: A sociocultural approach.* Routledge.

[35] Neil Mercer and Rupert Wegerif. 2002. Is 'exploratory talk' productive talk? In *Learning with computers*. Routledge, 93–115.

[36] Regina Nuzzo. 2014. Statistical errors. *Nature* 506, 7487 (2014), 150.

[37] W James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. (1999).

[38] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark, In 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. *arXiv preprint arXiv:2111.15366*.

[39] Johnny Saldaña. 2021. *The coding manual for qualitative researchers* (4 ed.). Sage.

[40] Jan-Willem Strijbos, Rob L Martens, Frans J Prins, and Wim MG Jochems. 2006. Content analysis: What are they talking about? *Computers & education* 46, 1 (2006), 29–48.

[41] Tim Van der Zee, Dan Davis, Nadira Saab, Bas Giesbers, Jasper Ginn, Frans Van Der Sluis, Fred Paas, and Wilfried Admiraal. 2018. Evaluating retrieval practice in a MOOC: how writing and reading summaries of videos affects student learning. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge*. ACM, 216–225.

[42] Rupert Wegerif. 2008. Reason and dialogue in education. *The transformation of learning. Advances in cultural-historical activity theory* (2008), 273–286.

[43] Michael Yeomans and Justin Reich. 2017. Planning prompts increase and forecast course completion in massive open online courses. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference*. ACM, 464–473.