The lack of generalisability in learning analytics research: why, how does it matter, and where to?

Mohammad Khalil
Centre for the Science of Learning & Technology (SLATE)
University of Bergen
Bergen, Norway
mohammad.khalil@uib.no

Paul Prinsloo
University of South Africa
Pretoria, South Africa
prinsp@unisa.ac.za

Abstract

Concerns about the lack of impact of learning analytics (LA) research has been part of the evolution of the field since its emergence as a research focus and practice in 2011. The preponderance of small-scale and exploratory nature of much of LA research are well-documented as contributing factors to the lack of generalisability, transferability, replicability and scalability. Through an analysis of 144 full research papers published in the conference proceedings of the Learning Analytics & Knowledge (LAK) Conference '22, 23 and 24, this paper provides an overview of the extent and contours of the lack of generalisability in LA research and pointers for making LA research more generalisable. The inductive and deductive analysis of the recent three LAK conferences provide evidence that a significant percentage (46%) of the corpus papers do not refer at all to generalisability or transferability, while few papers report on the scalability of their research findings. While the crisis of replicability/reproducibility is a wider concern in the broader context of research, considering and reporting on generalisability and transferability is integral to the scientific rigour. We conclude our paper with a range of pointers for addressing the lack of generalisability in LA research including, but not limited to expanding data, methodological adaptation and the potential of open science.

CCS Concepts

• **General and reference** → Document types; Surveys and overviews; • **Computing methodologies** → Artificial intelligence; Planning and scheduling; Planning with abstraction and generalization.

Keywords

generalisability, impact, learning analytics, replicability, transferability, scalability

ACM Reference Format:
Mohammad Khalil and Paul Prinsloo. 2025. The lack of generalisability in learning analytics research: why, how does it matter, and where to?. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706489

LAK 2025, March 03–07, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706489

1 Introduction

The potential of learning analytics (LA) to impact student learning and the contexts in which their learning occurs has been the central value proposition in LA since its emergence in 2011. From as early as 2012, [23] mentioned concerns about the impact of LA, and since then questions about the impact have been a regular feature in LA research (e.g. [8; 20; 16; 29]) In 2019, Dawson, Joksimovic, Poquet and Siemens stated that LA research "has failed to fully deliver on its promise to impact education and educational institutions" [8] (p. 446). Reasons for the lack of impact are many, such as "a continuing predominance of small-scale techno-centric exploratory studies that to date have not fully accounted for the multidisciplinary that comprises education." [10] (p. 446), and the complexity of education [11]. Research by [9] added another important consideration in measuring the impact of LA research and interventions by pointing to the fact that "the impact of support interventions can be explained in part, by the methodological constraints associated with these forms of research" (p. 477). [10] propose for LA to have impact, "there must be a purposeful shift to move from exploratory models to more holistic and integrative systems-level research" (p. 446).

Underlying and linking these concerns about the lack of impact, is the issue of generalisability, transferability, replicability and scalability. Of these four issues, the lack of replicability in research made headlines less than a decade ago in the context of psychological research [1]. The need that research findings should be replicable is one of the cornerstones of the scientific method [51], and the lack of replicability seriously impacts the credibility of research (see [17] for different typologies on replicability/reproducibility). Later in the paper we provide an overview of these terms and how we approached them in the context of LA research.

Our paper unfolds as follows: We firstly provide a brief problem statement, main research question, and value contribution before clarifying our understanding of generalisability, transferability, replicability and scalability as a baseline for the investigation into how these play a role in limiting the impact of LA research. This is followed by a literature review on the specific topic of generalisability in LA research. We then provide an overview of our methodology, followed by the analysis and discussion. The penultimate section presents a range of pointers on how to increase the generalisability of LA research before a section in which we address a number of limitations and conclude the paper.

1.1 Problem Statement, research question and value contribution

In the light of concerns about the lack of evidence of the (large-scale) impact of LA on the learning and institutional landscape, the lack of generalisability, transferability, replicability/reproducibility and scalability are as possible contributing factors.

This paper sets out to get a sense of . . .:

- to what extent LA research recognises a lack of generalisability;
- the contours of this lack of generalisability;
- pointers for making LA research more generalisable

The value contribution of this paper is found in its attention to generalisability in the context of LA research, how it is recognised and addressed in LA research to establish a tentative baseline for proposing a number of recommendations to increase the generalisability of LA research, as well as contribute to the impact of LA research on institutional strategies to support students' learning.

2 Clarification of Terms

In considering generalisability in the context of LA research, various other terms such as replicability, transferability, and scaling/scalability are at play. Though these terms seem to contain some overlapping elements and sometimes are used interchangeably, it is necessary to seek some conceptual clarity especially in the light of the different paradigms and understandings of scientific rigour informing qualitative, quantitative and mixed methods research [32; 39; 40]. It falls outside the scope of this paper to map the various issues in the "replication crisis" and suggested solutions such as, inter alia, the relative "low statistical power in single replication studies" [40] (p. 487). Also noteworthy are several reports on the failure of transferability of research findings (e.g., [49]).

While the criterion of generalisability, together with validity and reliability in quantitative research is well-established, generalisability in qualitative research is a contested terrain. Despite the fact that qualitative studies in LA research are relatively rare [13; 41], it is often presumed that generalisability does not apply to qualitative research [12]. There is however a different view that, "in qualitative domains, generalisability is possible provided that, first, generalisability is the main objective of the study; second, due precautions concerning the philosophy and terminology selected are taken" [3]. As Drisko [12] shows, in the context of qualitative research, generalisability should not be confused with transferability which is defined as "a process of abstraction used to apply information drawn from specific persons, settings, and eras to others that have not been directly studied" (p. 2). Despite this definition sounding like generalisability, Lincoln and Guba in 1985 [35] introduced the term 'transferability' in their "effort to distinguish the yield and use of interpretive qualitative research results from positivist conceptualizations aiming for law-like prediction and control" (in [12], p. 2). As such generalisation in quantitative research, or as [35] propose, 'nomic generalisations' are "assertions of enduring value that are context free" (p. 110). [12] further points out that generalisability is not the main purpose of qualitative research, and that transferability is a more appropriate term than generalisability, allowing other researchers to translate findings from one context to another. (See [3] for a view of the conditions under which theoretical generalisation in qualitative research is possible).

In quantitative research, generalisability and transferability are furthermore not mutually exclusive in the sense that the potential for generalisability depends on how transferable the findings are to produce the same effects in different circumstances or contexts, whether the same results can be achieved using different measurements, and whether the effects will be equally valid when applied to the whole population from which the sample was selected (e.g. [19; 48]).

Interesting to note that in much of the literature on replicability or reproducibility of research, no mention is made of transferability (e.g. [1]). Even when research findings may be replicable, *does not mean that they are generalisable or transferable.* The lack of replicability has been attributed, among other factors, to the pressure on the public, low statistical power or poor analysis, non-transparency regarding methods and availability of original data [1]. While replicability or reproducibility in qualitative research has never been the main objective, [53] proposes that replicability in qualitative research is possible under four conditions namely "the type of replication, the researcher's epistemological stance and approach, the aim of the replication, and the nature of the study" (p. 365). (Also see [44]).

Pertaining to scalability as a possible variable in the lack of impact of LA research on practice, it is important to distinguish between how small-scale studies impact on the generalisability of their findings, and the potential to scale findings or interventions. From this follows that small scale studies, with findings that are not generalisable, can still be scaled despite the fact that the findings may not be generalisable, or transferable to other contexts. There may also be studies that can be generalised and transferred but may not scale well.

Evident from the above cursory overview of terms such as generalisability, transferability, replicability and scalability is the fact that it falls outside the scope of this paper to resolve these issues in the context of LA research, and secondly in the context of this paper focusing on generalisability of LA research as possible variable in the lack of impact, we have to consider all of these, due to the possibility that these terms, although they may refer to different aspects of the lack of impact, may have been used interchangeably by authors.

Figure 1 attempts to illustrate the often-interchangeable way generalisability, transferability, replicability and scalability are used, but also the unique elements in each of these different terms.

3 Generalisability in LA Research: A Brief Literature Review

This brief literature review provides a *narrative* overview of literature on generalisability in LA research as a basis for the research questions and design for this paper.

One of the earliest attempts to map issues pertaining to the impact of LA is the research by [23] who refer to the role of stakeholders, the objective of LA, the nature of data, the choice of instruments as well as internal and external limitations. Impacting on generalisability is the fact that if "one of the parameters changes, the outcome and anticipated benefits will change" (p. 54). This

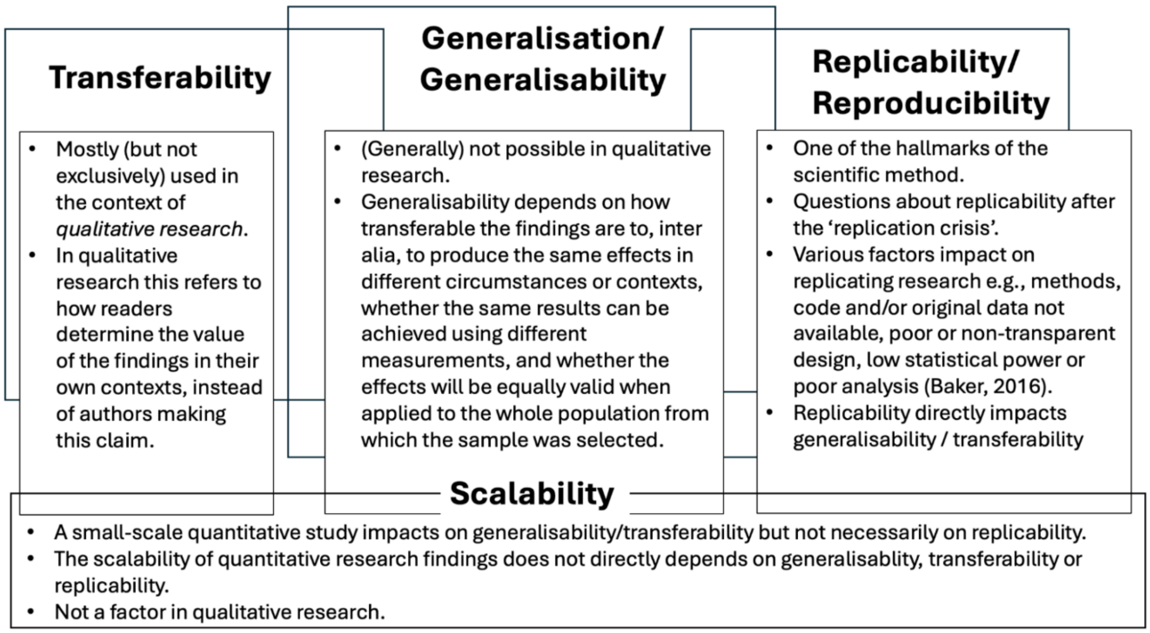

Figure 1: An overview of the linkages and differences between transferability, generalisability, replicability and scalability

aligns with concerns about the "dearth of empirical studies that have sought to evaluate the impact and transferability of this initial work across domains and contexts" [20] (p. 65). The lack of generalisability is also mentioned in 2017 in an article by Ferguson and Clow "'Where is the evidence?' A call to action for learning analytics" [16]. They point to "Limited attention to validity, reliability and generalisability" and they propose that LAK conference organisers should specifically address issues pertaining to validity, reliability and generalisability in the paper review process. (Also see [42]).

In their exploration of the possible reasons for the lack of impact and generalisability of LA research on educational systems, Dawson et al. [10] start by mapping the maturity of LA taking cognisance that it is a "bricolage field, emerging from the convergence of multiple disciplines" (p. 447). The multidisciplinary origins and nature of LA and the different epistemologies, ontologies, methodologies and theories in this nexus of multiple disciplines "generates many benefits but it is also accompanied with many challenges" (p. 447). The authors found that most of LA research can be classified as "exploratory research, focusing on students enrolled in independent courses or small projects" (p. 452) and that "few studies to date that have attempted to extend exploratory works *to evaluate the results or suggested actions* to interrogate the proposed model and feedback to theory" (p.452, italics added). The authors state that LA "remains mired within a phase of exploratory analyses." Interesting, in mapping possible reasons for this state of affairs, the authors mention the impact of external factors such as "at numerous levels such as politics, funding sources from international, national and local sources, availability and access to data, opportunities and limitations of methods and access to expertise through our own collaborative networks" (p. 453). Also contributing to the preponderance of exploratory models is the entanglement of education with a variety of broader systems and networks adding to the complexity of identifying appropriate interventions. In this regard LA research aligns with concerns in social science research regarding the replication and validation of research findings. A further complicating factor impacting on the generalisability of LA research is the fact that the most frequent research strategy is case study research that are often small-scale contextual studies that are not generalisable [13]. This study also reports an absence of large-scale and longitudinal research.

Linking the efficacy of LA interventions with generalisability, [33] recommend more "scientifically driven learning analytics" to provide a sound evidence base "for the feasibility, effectiveness, and generalisability of such interventions" (p. 1) as there are "there is very little evidence for the generalisability of these effects" (p. 25). They conclude that generalisability is and should be a central concern for future LA research as "replicating past results in new contexts may be the best way forward in validating LA intervention programs" (p. 30). These findings and recommendations are echoed in the systematic review by [14] who found that the majority of models for predicting student success *cannot be generalised* due to design issues (e.g., utilising aggregated behaviours at the end of the course, differentiating between important factors and predictive factors, and the specific requirements of courses). The authors propose "Future research should focus on providing actual early warning predictions and proposing transformation rules for improving a model's generalisability" (p. 59). The systematic review furthermore reports on the 29.4% (from a total of 265) of articles reported on a proof of concept with small scale of data analysis. The generalisability of LA research is also impacted by contextual differences such as geopolitical and institutional contexts and methodological choices [27] as well as “the relevance of particular measurements [which] operates in a matrix of valuations that instil certain information as more useful given certain theoretical aims” [46] (p. 274). LA researchers cannot and should not assume that their findings will generalise across contexts and methodological choices. Interestingly, the authors conclude that “If the methodologies employed by learning analytics researchers are to take a more qualitative turn, however, such approaches open possibilities of transforming notions of validity, reliability, and even generalisability to more situated and mediated notions” (p.282).

The importance of contextual, methodological and research design choices on generalisability in LA research is further illustrated by [57] who refer to methodological challenges impacting on generalisability in the context of the Multi-Modal Learning Analytics (MMLA) and they remark that “Most of the predictive analytics studies (71%) were conducted with a small-scale samples (n<50), which could undermine the generalisability of the models to other samples” (p.11). Possibly more concerning is their findings that “Two-thirds of the reviewed studies are not replicable as they have not disclosed sufficient methodological information regarding the sensors or analysis techniques used in the study” (p. 11). In the section, “Implications for practice,” [57] express their concerns about the reporting standards in MMLA research and they propose “studies to include enough methodological details,” emphasising that such details “are essential for MMLA research to be less vulnerable to the replication crisis that has been found across multiple scientific research fields” (p. 12).

As illustrated in Figure 1, the issue of generalisability in LA research is also linked to claims about replicability/reproducibility. Two articles that specifically address the issue of replicability/reproducibility of LA research are the papers by Kitto et al. [28] and Haim et al [25]. In the first paper, [28] introduces the problem of replicability of research by stating that research that is considered statistically significant in many different fields “cannot be replicated in new contexts and a dramatic restructuring of what counts as validated theory is now underway” (p. 303). In the context of LA research, attempts to replicate research results in new contexts “fail more often they succeed” (p. 304). Though there are many reasons for the failure to replicate LA research findings, [28] mention that “the different choices made in performing a data analysis create many more degrees of freedom than are normally assumed” resulting in the significance estimates being “overly optimistic” (p. 304). In response, the *Journal of Learning Analytics*, for example, attempted to address the issue of replicability through the LAK data challenge and the Pittsburgh data shop where open access to educational data was provided [28].

The replicability/reproducibility of LA research are furthermore influenced by a range of LA procedures, choices with regard to coding and analysis of the results that are not well documented and communicated in the subsequent publishing of the results. Commenting about the choices in LA research that impact on replicability/reproducibility, [28] refer to how national and institutional policies and systems shape data collection e.g., the type of data collected. Other than these systematic influences, we should also consider the ways data are collected, how the subsets of these collected data are being defined, and specifics regarding the student cohort which all add to the complexity of the phenomenon LA research is trying to understand [28]. This aligns with the work of [38], who framed the implementation of LA as a 'wicked problem' due to the inherent complexity of higher education systems (p. 17), see also [10].

Linked to the research by [28] is the paper by [25] proposing a number of principles to guide reproducibility in LA research. Central to their proposal is the notion and practices of open science, and how it impacts on the reproducibility of LA research. They found that in the analysed corpus, “less than half of the research adopted standard open science principles” and they were “unable to reproduce any of the papers successfully in the given time period” allowed for in their methodology (p. 156). One of their recommendations with which they conclude their paper is to suggest that open science and reproducibility designs should be promoted by conference organisers encouraging researchers to share processes, methodologies and associated data.

As Figure 1 illustrates, generalisability, transferability, and replicability/reproducibility are also linked to the notion of scalability. [31] argue that though the value proposition of LA research is well-established, “the kind of intended impact is not always clear.” For example, the features of the contexts in which learning occurs are specific to particular sites and “make achieving impact challenging.” There are implicit tensions between scaling LA interventions to reach as many students as possible or to “scale impact on learning” (comparing learning gains with other interventions). The authors “contend that while certainly it is possible to achieve scale in learning gains while also scaling populations, a focus on the latter can lead to a disconnect between learning analytics implementations and impact of scale on student learning.” (Also see [57]).

The above literature review describes and briefly analyses published literature on the issue of generalisability of LA research. The literature review can be summarised as follows:

- Central to LA research since its emergence as a research focus and practice was the question of impact [8; 20; 23].
- The proposal by [16] to pay more explicit attention to validity, reliability and generalisability in LA research may have gone unheeded.
- The inherent nature of LA research as a “bricolage field, emerging from the convergence of multiple disciplines” continues to shape not only LA research but also influence the exploratory nature of much of LA research [10]. As such LA research shares concerns in the social sciences where issues of replication and validation of research findings remain a concern.
- There are many variables impacting on the lack of generalisability of LA research such as, but not limited to a lack of explicit focus on generalisability of LA research from the initiation and design of the research [33], the small scale of data analysis [13; 14], the lack of methodological details in how LA research are reported [28; 57], the differences between continents, application contexts and methodological choices [27], the potential of adding qualitative data analysis to the design of LA research [46], the potential of open science principles and practices [25] and the lack of longitudinal studies [13].

Having established not only the complexity of finding evidence of the impact of LA research as well as the concerns about generalisability and replicability in LA research as one of the possible reasons for the lack of impact, but we also map the research design and methodology in the next section, followed by the analysis and findings, and pointers for consideration.

4 Methodology

In line with the purpose of this article to contribute to not only understanding how LA research approaches and presents generalisability, but also to inform and enhance the generalisability of future LA research, we took inspiration from [10] methodology and analysed 144 full research papers from the three most recent LAK conferences – LAK '22, 23 and 24. [See the Limitations section for an acknowledgement of the impact of the scope of the selection]. Following the coding process used by [28] (p. 306) the two authors discussed (1) how each one would approach the deductive and inductive analysis; (2) in what order would we code; (3) how we defined/would approach each of the terms central to the analysis; (4) how we would differentiate between closely related terms such as generalisability and transferability; and (5) how we would approach a coding when we were not sure. In the case of the latter, the authors decided to colour code such instances with bright pink in order to resolve through discussions. We have also made available the working sheets in this link: https://kortlink.dk/2qx3t

The full papers were downloaded and deductively and inductively coded using the following coding heuristic: (1) Limitation is mentioned in the literature review; (2) limitation is mentioned in the discussion of the results; (3) the research design (quantitative, qualitative, mixed); (4) limitations due to the small scale of the study; (5) limitations in terms of generalisability; limitations pertaining to transferability; (7) limitation in terms of tools/methods; (7) limitation mentioned in terms of replicability; (8) Limitations - other; and (9) solutions mentioned to address the limitations or lack of generalisability.

The coding heuristic was confirmed after each of the authors coded 20 papers each, discussed the usefulness of the coding heuristic, and consensus about the range of the aspects covered in the coding heuristic. After that, each author reviewed half of the papers. This process was repeated after another 20 papers were coded, the heuristic reviewed, and consensus established. Regarding the issue of inter-rater reliability, (IRR) we took cognisance of the issues raised by [28] presentation of the different ways to calculate IRR. In recognition of the issues raised, in addition to using Fleiss' kappa to assess agreement between the authors, we also added as much appropriate data as possible to ensure transparency in the process. According to [18], a kappa value above 0.81 indicates strong agreement, 0.61–0.80 signifies good agreement, 0.41–0.60 reflects fair consensus, and values below 0.41 suggest poor agreement. The final kappa when comparing the results of the coding authors indicated a good level of agreement ($\kappa = 0.76$, subjects =154, raters = 2, and $p < 0.005$) with low measure of controversy.

We *excluded* from the downloaded corpus:

- Systematic, scoping and critical reviews. Where such a paper addressed generalisability and the other related terms, we included these in the Literature review section.
- Conceptual research unless specifically addressing generalisability and the other terms.
- Editorials, short and practitioner papers and workshops from the analysis.

The corpus for analysis included **154** of articles from LAK '24, LAK '23, and LAK '22, of which **10** papers were excluded due to these being systematic, scoping and critical reviews or conceptual research, resulting in **144** articles in the final corpus for analysis.

The deductive and inductive analysis resulted in two distinct but linked results. *Firstly*, we developed a heuristic of five (5) categories and classified the articles according to the combination of mentioning/discussing of limitations, generalisability, replicability, reproducibility, transferability, and/or scale/scalability (see Table 1 below). To ensure rigour and transparency regarding the categorisation process, one researcher categorised all the papers from LAK '22, 23, and 24 while the second reviewer verified the categorisation. The inter-rater reliability for the categorisation was ($\kappa = 0.93$, subjects =66, raters = 2, and $p < 0.005$).

The *second* part of the analysis involved an inductive coding of the limitations' sections of articles in the corpus in order to develop a thematic analysis of pointers for increasing the generalisability of LA research. Once again, the coding and theming process was informed by [28]. Both researchers coded the limitations' section with specific attention to how authors suggested addressing the limitations' sections, after which the differences in coding were discussed and resolved and collectively themed.

It is important to note that the starting point of the categorisation was whether the papers mentioned any *limitations*, either as a distinct section or in the discussion of the findings and analysis. This allowed us to treat quantitative, qualitative and mixed research designs equally as a starting point.

Two issues are worth mentioning, firstly the fact that generalisability is not a criterion for establishing rigour in qualitative research, but qualitative research adheres to different criteria for scientific rigour such as establishing trustworthiness and recognising the limitations inherent in transferring the findings from qualitative research to other contexts, populations, etc. (e.g., [50]). We therefore looked for both terms, regardless of the research design - generalisability and transferability. A second issue added a layer of complexity to the analysis - namely the fact that many papers did not mention the selection of a specific research design such as quantitative, qualitative or mixed methods. While the non-declaration of a specific research design is not a central issue in this paper, it has implications for how LA research is reported and what scientific transparency is required from authors. If authors do not explicitly declare their choice of a research design, how did they approach the provision of evidence for ensuring validity, reliability and generalisability (in quantitative research), and trustworthiness and transferability in qualitative research?

5 Analysis and Discussion

5.1 Categorisation of limitation, generalisability and related terms

Table 1 provides an overview of the 144 papers analysed to determine their categorisation.

Table 1: Overview of categories and frequencies

Category	Description	Number of papers in this theme
Category 1	**No limitations mentioned or acknowledged** but **DOES NOT MENTION OR REFER TO** generalisability AND/OR replicability AND/OR reproducibility AND/OR transferability, AND/OR reference to scale/scalability	19
Category 2	**No limitations mentioned or acknowledged BUT SOME details** or references to generalisability AND/OR replicability AND/OR reproducibility AND/OR transferability, AND/OR reference to scale/scalability	10
Category 3	**Limitations acknowledged/ mentioned** but **DOES NOT MENTION OR REFER TO** generalisability AND/OR replicability AND/OR reproducibility AND/OR transferability, **BUT NO** reference to scale/scalability	47
Category 4	**Limitations acknowledged/ mentioned AND REFERS TO** generalisability AND/OR replicability AND/OR reproducibility AND/OR transferability, **BUT NO** reference to scale/scalability	64
Category 5	**Limitations acknowledged/ mentioned AND REFERS TO** generalisability AND/OR replicability AND/OR reproducibility AND/OR transferability, **AND refer** to scale/scalability	4
	Total	144

As can be seen in the above table (Table 1), **44%** (n= 64) of the papers were found to fall in **Category 4** where limitations were explicitly acknowledged and generalisability AND/OR replicability AND/OR reproducibility AND/OR transferability, but with no reference to scale/scalability. The second-most populated category was **Category 3** with **33%** (n= 47) of the papers acknowledging limitations but does not refer to generalisability and/or replicability and/or reproducibility and/or transferability, as well as no reference to scale/scalability.

Significant is the fact that a total of 20% (n= 29; **Category 1** and **Category 2**) of the papers *do not* acknowledge *any* limitations in the paper while only four (n= 4) papers acknowledging limitations and referring to generalisability and/or replicability and/or reproducibility and/or transferability, as well as to scale/scalability (**Category 5**).

An example of an article that falls in **Category 1** (no limitations mentioned and no mention of generalisability and/or replicability and/or reproducibility and/or transferability, as well as no reference to scale/scalability) is the article by [43] in which they report on "The effects of learning analytics hint system in supporting students problem-solving". Their research found that "... the LA hint system improved students' problem-solving SE significantly" (p. 83). (Also see [60]).

An example of a **Category 2** article which does not refer to any limitations but with a reference or discussion on generalisability and/or replicability and/or reproducibility and/or transferability and/or scale/scalability is the article by Dwivedi et al. [15] who reports on "the 'Effecti-Net' architecture, a sophisticated deep learning model that integrates data from multiple sensor modalities, including Electroencephalogram (EEG), Eye Tracker, Galvanic Skin Response (GSR), and Photoplethysmography (PPG)" which "offers educators and content creators a comprehensive framework that promotes the development of more engaging educational content" (p. 667). [15] state "By integrating information from multiple sources, the model is more robust to noise and individual differences, *resulting in better generalisation*" (p. 674; emphasis added).

The article by [7] "Protected attributes tell us who, behaviour tells us how: A comparison of demographic and behavioural oversampling for fair student success modelling" is an example of **Category 3** where limitations are acknowledged but where there is no discussion or references to generalisability and/or replicability and/or reproducibility and/or transferability, as well as no reference to scale/scalability. Chang et al. [4], in their article "Towards an automatic approach for assessing program competencies" refer in their literature review to several limitations in various research designs and methodologies preventing "valid comparisons across institutions and studies" (p. 120). After presenting their research findings, they have a substantive section on the limitations in their own research, without explicitly mentioning generalisability.

An example of **Category 4** (limitations are acknowledged with some discussion or references to generalisability and/or replicability and/or reproducibility and/or transferability, but no reference to scale/scalability) is the article by [24] who included a Limitations section and specifically mention "This research should be replicated using different content areas across different subjects to fully understand the heterogeneity of the impact the problem types can have on student learning. A further limitation of the current study

Table 2: Themes extracted for proposed pointers addressing generalizability in LA research. ***Some papers appear on multiple themes.***

Theme	Description	Number of papers in this theme
Data expansion	Solutions involving increasing sample sizes, incorporating data from additional sources, or using more diverse datasets.	17
Methodological adaptation	Extending or adapting methods, models, or algorithms to different contexts or types of data.	12
Cross-validation and domain transfer	Proposals to replicate studies across different settings, cohorts, domain, or datasets to validate findings including theoretical validation	12
Multimodal integration	Combining multiple types of data (e.g., observational, self-report, audio, video, sensors) to improve robustness.	9
Advanced analytics	Applying advanced or new analytical methods (e.g., AI models, Explainable AI).	8
Bias and fairness	Addressing fairness and bias in models or data to enhance generalisability.	6
Augmentation and synthesis	Using techniques like data augmentation or synthetic data generation.	4
Open science and reproducibility	Encouraging practices such as open data, replication studies, or improving reporting standards.	3

is the lack of student demographic information" (p. 516). Cloude, Baker and Fouh [6] in their article "Online help-seeking occurring in multiple computer-mediated conversations affects grades in an introductory programming course" state "To identify the extent to which the results of our study could be generalised, certain limitations need to be acknowledged" (p. 386). Also see [55] who specifically mention that "A key challenge in building predictive analytics is model generalisation" (p. 713). After discussing their results, the authors state ""we cannot guarantee that the regression estimations from our analysis can be generalised well to other institutions" (p.723).

There were only four (4) papers in **Category 5** acknowledging limitations and refer to generalisability and/or replicability and/or reproducibility and/or transferability, as well as scale/scalability. An example of this category is the article by [61] who specifically mention a number of limitations in their article such as "our procedures for extracting and coding the dialogue content were not fully automated, which limited the initial scalability of our approach (p. 194). They also recognise that the size of their participating teams limited the findings' generalisability. "While the effects reported in this study were strong enough to obtain statistical significance, we do not assume that our results necessarily generalise beyond the specific healthcare context we examined. Nonetheless, our results may be insightful for future studies of similar contexts. (p. 195). Also in this category, the article by [2] said that "This work is the first to predict course load *at scale*, generalising to over 10,000 courses at a large public institution and going beyond time load considerations by incorporating more holistic measures such as mental effort and psychological stress" (p. 219; emphasis added).

Interestingly, none of the four articles in this category *claimed scalability*, though reporting on large-scale research. See, for example, [34] who "conducted a study on a large-scale dataset (5,165 students and 116 qualified tutors in 18,203 online tutoring sessions) of both effective and ineffective human-human online tutorial dialogues" (p. 282) but though the research used a large-scale data, the authors acknowledge a range of limitations that defies generalisability.

5.2 Themes and pointers for addressing the generalisability of LA research

In our analysis of the strategies proposed to address the recognised limitations of generalisability in LA research, we found that 61 out of the 144 reviewed studies explicitly offered solutions aimed at broadening the scope and applicability of their findings to LA research.

To provide a clearer understanding of the proposed solutions, we identified and classified them into 8 distinct themes, as summarised in Table 2. Each theme represents a unique aspect towards more generalisable LA. We also report the number of papers corresponding to each of the themes. Details and examples follow.

5.2.1 Data expansion. Data expansion refers to strategies for increasing the breadth, depth, and diversity of data used in LA. The general principle is that broader and more varied datasets lead to more robust, scalable, as well as generalisable findings. This theme preserves the largest share of pointers towards generalisability in LA.

In [58], the authors argue that to better understand the relationship between learning features, such as topic areas and student knowledge, there should be deeper studies that consider how these elements interact. The authors proposed that in order to generalise their experiment, they argue for larger sample sizes and more diverse data to recognise patterns that may not emerge in smaller or more homogeneous samples.

In [26], the authors discussed that expanding the dataset with additional labelled samples can be critical for their GPT modelling. By involving more human-labelled data, the dataset is expected to be richer and enable more nuanced insights. However, labelling samples for further machine learning training can be both time-consuming as well as expensive [54]. On the other hand, [45] highlight the need for data from private universities in addition to public institutions in Brazil to enhance further adoption of LA. By expanding the dataset to include private universities, the researchers would gain a more holistic understanding of LA adoption in the country which makes their findings more approachable by the decision makers at the country level.

5.2.2 Methodological adaptation. Twelve papers reported extending methods, models, and algorithms can support replicable and generalisable findings from LA. In [5], the authors propose the need to develop an adaptable Application Programming Interface (API) to a broader customisable parameters of a generative AI writing support tool in order to make the system more versatile to other human-AI writing applications. Another example of methodological adaptation is the approach proposed by [46], which proposes to integrate qualitative research alongside data-driven methods. While this will enhance the generalisability of LA research as discussed by [46], it may also introduce additional complexity for LA practitioners.

5.2.3 Cross-validation and domain transfer. Cross-validation and domain transfer have been explored in 12 out of the 61 studies. This theme entails that the findings of one context or dataset hold true in other contexts, for example interventions to a different group of learners. From the corpus, [47] underscore the necessity of validating their results regarding essay writing to deepen understanding of metacognitive processes. They suggested carrying out analysis of temporal and sequential patterns to strengthen duration-based inferences. [47] highlight how language proficiency may also shape learner outcomes, which was reported as a limitation for their sample of participants that vary in language proficiency levels. This approach to cross-validation would support the broader generalisation of their findings.

5.2.4 Multimodal integration. The next theme, multimodal integration, involves using multiple data sources to enhance the robustness of research. In classroom settings, this approach can improve the variety of data and provide a more comprehensive understanding of learning environments. For instance, although [56] collected over 173 million data points to gain detailed insights into teachers' spatial behaviours, they noted that combining spatial and audio data can be advantageous for making well-rounded judgments. The absence of audio sensors limited their ability to capture verbal communication which by then highlighted the possible need for integrating audio data to complement spatial information.

5.2.5 Advanced analytics. Some LA researchers have recognised the value of conducting additional analyses and apply new AI-powered analytics methods to address challenges related to generalisability. A recent study by [52] exemplifies this approach using Large Language Models (LLMs) to identify student challenges and improve the fit of personalised interventions. The discourse presented by the authors aligns with the view that LLMs have "high applicability to perform a task even in a specific domain and have a higher capability to generalise into other contexts than the rule-based/supervised ML" (p. 482).

5.2.6 Bias and fairness. Bias-free and fairness modelling can broaden the applicability of research in LA by ensuring that findings are valid across diverse contexts and populations [7], [55]. Fairness and bias-free models address systematic errors and prevent skewed outcomes, thereby promoting generalisability of the results. For instance, [59] have introduced Bayesian Knowledge Tracing (BKT), while [55] have developed the concept of 'fairness shift'. Both approaches contribute to creating more transferable and portable machine learning models.

5.2.7 Augmentation and synthesis. Several researchers in the field have introduced methods for augmenting original datasets using synthetic methods to create additional scenarios for LA experiments. For instance, [30] and [37] employed deep learning techniques to generate synthetic data and demonstrated that their approach can enhance both the scalability and generalisability of LA models. Building on related aspects of data synthesis, a recent research study has explored methods to balance privacy and fairness during the generation of synthetic datasets [36].

5.2.8 Open science and reproducibility. Institutions can contribute to the generalisability of LA via open science practices. In our mapping, three key studies were identified as actively contributing to this discourse. For instance, [25] emphasise the importance of replicating research and fostering collaboration with original authors, stating that "it would be useful to replicate this work and contact the authors so that any previously unclear or uncertain principles can be specified and provided" (p. 162). Similarly, [22] recommend promoting reproducibility by openly sharing analysis scripts, noting that such practices "encourage other universities to replicate the analyses with their data" (p. 495).

5.3 Limitations

Our paper on examining the lack of generalisability in LA has certain limitations. In this paper we did not assess whether authors of quantitative studies addressed reliability and validity in their reporting of research designs and findings. Together with generalisability, validity and reliability form the basis for evaluating scientific rigour in quantitative research, in the same way that trustworthiness functions in qualitative research. While we recognise the limitations of only analysing all the full papers (144) in three years of LAK conferences, the categorisation was developed in the analysis of papers from LAK '23 and LAK '24 and tested and applied in the analysis of full papers from LAK '22 We assumed that the last three years' proceedings from the LAK conference reflect the maturity of the field. We also acknowledge that some short papers may have offered insights into generalisability in LA research. Lastly, while we aimed to cover the key concepts under the umbrella of generalisability including transferability, replicability, and scaling/scalability, we may have overlooked other relevant concepts.

These limitations may affect our findings, yet highlight the need for further exploration on this key topic in future research.

6 Pointers and Conclusions

Based on the categorisation of the corpus of analysis, as well as the thematic representation of proposals to address the limitations and generalisability of LA research, we propose a number of pointers to advance discussions about generalisability of LA research and increase the scientific rigour and reporting of research findings.

1. It is crucial to make the reporting of limitations of any LA research a non-negotiable element in the presentation of LA research. A significant proportion of LA research in this study (20%) did not include any discussion of the limitations of the reported research. Conference organisers and journal editors could make this an integral part of the reporting of findings, equally important to declarations of ethical clearance, funding, and open data.
2. The analysis showed that papers often did not explicitly mention whether the research paradigm was quantitative, qualitative or mixed methods. While researchers may assume that the paradigm is implicit in the reporting of the design and methods of data collection and analysis, the absence of an explicit mention of the research paradigm may point to a possible disregard to acknowledge and address requirements of scientific rigour in the reporting of the findings as proposed by [16]. This can limit the replicability of research due to ambiguity in the methodological approach. Transparency regarding methodological and analysis choices have been mentioned numerous times in the literature as an important factor that can increase the generalisability but also the rigor of LA research (e.g., [28]). Reviewers and conference organisers should encourage, if not sanction, an explicit consideration of how the methodological and analytical choices impact on the generalisability of the reported findings.
3. The analysis confirms that there is still a predominance of "small-scale techno-centric exploratory studies" and exploratory models [10] as well as case study research designs [13] in LA research. There is also a lack of longitudinal The dominance of exploratory research may be due to a variety of factors such as, inter alia, a realisation of the complexity of understanding student learning in increasingly open, recursive and dynamic systems, the difficulty of multidisciplinary research and the lack of longitudinal research (e.g., [13]). Without ignoring the importance of any form of LA research, conference organisers and journal editors can steer LA research by encouraging and specifically inviting larger scale and longitudinal research. For example, the Society of Learning Analytics Research (SoLAR) can offer funding support for large-scale and longitudinal LA research.
4. There are currently very few qualitative studies in LA research and [46] recommended that qualitative research can "open possibilities of transforming notions of validity, reliability, and even generalisability to more situated and mediated notions" (p.282). Considering that LAK conferences and special issues in *the Journal of Learning Analytics* often steer the focus of LA research, there is much potential to enrich LA research and understandings of generalisability and transferability through embracing qualitative and multidisciplinary approaches to understanding students' learning and the contexts in which their learning occurs. Qualitative and mixed methods' research may also fall outside of the expertise of review panels resulting in such research, despite their rigor, are not accepted. Conference organisers and journal editors should ensure that their review panels include researchers well-versed in qualitative and mixed methods' research.
5. Almost half of the papers in this analysis (46%) does not refer at all to generalisability and/or transferability (categories one and three, Table 1). Considering that generalisability (in quantitative research) and transferability (in qualitative research) are essential elements in the establishing scientific rigour, it is clear that the recommendation of [33] that "generalisability of LA interventions should be a fundamental focus for future study" (p. 30) and others (e.g. [20]; [28]) remains unheeded. Also see Recommendation 2.
6. Only 4 studies in this analysis referred to issues of scale and/or the scaling of the findings. This seems to indicate that the concerns expressed by [21] about the predominance of small-scale and exploratory studies in LA research remains a concern. Also see Recommendation 3.
7. The themes presented in Table 2 and the discussion of the themes provide a number of useful pointers to address the generalisability of LA research ranging from, but not limited to data expansion, methodological expansion and the potential of open science.

We conclude by reaffirming that the issue of generalisability in LA research remains both critical and persistent. Many LA studies rely on small-scale, post-hoc designs that, while yielding promising results, often lack broader applicability. This limitation aligns with prior discussions emphasising that LA should avoid endorsing a one-size-fits-all approach [21]. While it is true that the findings of one context cannot always seamlessly be applied to another, this argument should not hinder efforts to extend the scope of LA research. With our mapping of the documented limitations mentioned in 144 papers, it is essential for the field to move beyond the boundary of small-scale studies and explore new ways to enhance the transferability and replicability of LA methods. Doing so would strengthen the field's impact, both within the LA community and across adjacent areas. In fact, addressing generalisability issues will contribute to the credibility and utility of LA research work.

Acknowledgments

The authors would like to express their appreciation to the reviewers and the meta-reviewer for their insightful and critical comments and suggestions.

References

[1] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604): 452-454.
[2] Borchers, C., & Pardos, Z. A. (2023). Insights into undergraduate pathways using course load analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 219-229).
[3] Carminati, L. (2018). Generalisability in qualitative research: A tale of two traditions. *Qualitative health research*, *28*(13), 2094-2101.
[4] Chang, X., Wang, B., & Hui, B. (2022). Towards an automatic approach for assessing program competencies. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 119-129).
[5] Cheng, Y., Lyons, K., Chen, G., Gašević, D., & Swiecki, Z. (2024). Evidence-centered Assessment for Writing with Generative AI. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 178-188).
[6] Cloude, E. B., Baker, R. S., & Fouh, E. (2023). Online help-seeking occurring in multiple computer-mediated conversations affects grades in an introductory programming course. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 378-387).
[7] Cock, J. M., Bilal, M., Davis, R., Marras, M., & Kaser, T. (2023). Protected attributes tell us who, behavior tells us how: A comparison of demographic and behavioral oversampling for fair student success modeling. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 488-498).
[8] Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 231-240).
[9] Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017). From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 474-478).
[10] Dawson, S., Joksimovic, S., Poquet, O., & and Siemens, G. (2019). Increasing the impact of learning analytics. In *Proceedings of the 9th international conference on learning analytics and knowledge*. 446–455.
[11] Dawson, S., Pardo, A., Salehian Kia, F., & Panadero, E. (2023). An Integrated Model of Feedback and Assessment: From fine grained to holistic programmatic review. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 579-584).
[12] Drisko, J. W. (2024). Transferability and Generalisation in Qualitative Research. *Research on Social Work Practice*, 10497315241256560.
[13] Drugova, E., Zhuravleva, I., Zakharova, U., & Latipov, A. (2024). Learning analytics driven improvements in learning design in higher education: A systematic literature review. *Journal of Computer Assisted Learning*, *40*(2), 510-524.
[14] Du, X., Yang, J., Shelton, B. E., Hung, J. L., & Zhang, M. (2021). A systematic meta-review and analysis of learning analytics research. *Behaviour & information technology*, *40*(1), 49-62.
[15] Dwivedi, D., Garg, R., Baghel, S., Thareja, R., Kulshrestha, R., Mohania, M., & Shukla, J. (2024). Effecti-Net: A Multimodal Framework and Database for Educational Content Effectiveness Analysis. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 667-677).
[16] Ferguson, R., & Clow, D. (2017). Where is the evidence? A call to action for learning analytics. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 56-65).
[17] Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/entries/scientific-reproducibility/~
[18] Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley & sons.
[19] Fyans, Leslie J. (Ed.). (1983). Generalisability theory: Inferences and practical applications. San Francisco: Jossey-Bass.
[20] Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*, 64-71.
[21] Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education, 28*, 68-84.
[22] Glandorf, D., Lee, H. R., Orona, G. A., Pumptow, M., Yu, R., & Fischer, C. (2024). Temporal and Between-Group Variability in College Dropout Prediction. I*n Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 486-497).
[23] Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, *15*(3), 42-57.
[24] Gurung, A., Vanacore, K., Mcreynolds, A. A., Ostrow, K. S., Worden, E., Sales, A. C., & Heffernan, N. T. (2024). Multiple Choice vs. Fill-In Problems: The Trade-off Between Scalability and Learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 507-517).
[25] Haim, A., Shaw, S., & Heffernan, N. (2023). How to open science: A principle and reproducibility review of the learning analytics and knowledge conference. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 156-164).
[26] Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., ... & Ker, C. L. (2024). Prompt-based and Fine-tuned GPT Models for Context-Dependent and-Independent Deductive Coding in Social Annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 518-528).
[27] Karumbaiah, S., Baker, R., Tao, Y., & Liu, Z. (2022, March). How does students' affect in virtual learning relate to their outcomes? A systematic review challenging the positive-negative dichotomy. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 24-33).
[28] Kitto, K., Manly, C. A., Ferguson, R., & Poquet, O. (2023). Towards more replicable content analysis for learning analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 303-314).
[29] Khalil, M., Prinsloo, P., & Slade, S. (2022). A comparison of learning analytics frameworks: A systematic review. In LAK22: 12th international learning analytics and knowledge conference (pp. 152-163).
[30] Khalil, M., Vadiee, F., Shakya, R., and Liu, Q. (2025). Creating Artificial Students that Never Existed: Leveraging Large Language Models and CTGANs for Synthetic Data Generation. In *Proceedings of the 15th Learning Analytics and Knowledge Conference (LAK'25)*. ACM. https://doi.org/10.1145/3706468.
[31] Knight, S., Gibson, A., & Shibani, A. (2020). Implementing learning analytics for learning impact: Taking tools to task. *The Internet and Higher Education*, *45*, 100729.
[32] Krauss, S. E. (2005). Research paradigms and meaning making: A primer. *The qualitative report*, *10*(4), 758-770.
[33] Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594-2618.
[34] Lin, J., Rakovic, M., Lang, D., Gasevic, D., & Chen, G. (2022). Exploring the politeness of instructional strategies from human-human online tutoring dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 282-293).
[35] Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Sage.
[36] Liu, Q., Deho, O., Vadiee, F., Khalil, M., Joksimovic, S., and Siemens, G. (2025). Can Synthetic Data Be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK'25).* ACM. https://doi.org/10.1145/3706468.3706546
[37] Liu, Q., Khalil, M., Jovanovic, J., & Shakya, R. (2024). Scaling While Privacy Preserving: A Comprehensive Synthetic Tabular Data Generation and Evaluation in Learning Analytics. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 620-631).
[38] Macfadyen, L. P., Dawson, S., Pardo, A., & Gašević, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, *9*, 17-28.
[39] Maxwell, J. A. (2021). Why qualitative methods are necessary for generalization. *Qualitative Psychology*, *8*(1), 111.
[40] Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.
[41] Moon, J., Lee, D., Choi, G. W., Seo, J., Do, J., & Lim, T. (2023). Learning analytics in seamless learning environments: A systematic review. *Interactive Learning Environments*, 1-18.
[42] Ochoa, X., Knight, S., & Wise, A. F. (2020). Learning analytics impact: Critical conversations on relevance and social responsibility. *Journal of Learning Analytics*, *7*(3), 1-5.
[43] Pan, Z., & Liu, M. (2022). The effects of learning analytics hint system in supporting students problem-solving. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 77-86).
[44] Peels, R., & Bouter, L. (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications, 4*(1).
[45] Pontual Falcão, T., Lins Rodrigues, R., Cechinel, C., Dermeval, D., Harada Teixeira de Oliveira, E., Gasparini, I., ... & Ferreira Mello, R. (2022). A penny for your thoughts: students and instructors' expectations about learning analytics in Brazil. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 186-196).
[46] Quick, J. D., Motz, B., & Morrone, A. (2023). Lost in Translation: Determining the Generalisability of Temporal Models across Course Contexts. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 273-283).
[47] Rakovic, M., Fan, Y., Van Der Graaf, J., Singh, S., Kilgour, J., Lim, L., ... & Gasevic, D. (2022). Using learner trace data to understand metacognitive processes in writing from multiple sources. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 130-141).
[48] SemanticScholar. (2019). Generalisability and Transferability. Retrieved from https://www.semanticscholar.org/paper/Generalizability-and-Transferability/40e3659649957a0e0183f743d576f18e4cbf4cd4
[49] Schloemer, T., & Schröder-Bäck, P. (2018). Criteria for evaluating transferability of health interventions: a systematic review and thematic synthesis. *Implementation Science*, *13*, 1-17.

[50] Smith, B. (2018). Generalisability in qualitative research: Misunderstandings, opportunities and recommendations for the sport and exercise sciences. Qualitative research in sport, *exercise and health, 10*(1), 137-149.

[51] Staddon, J. (2017). Scientific method: How science works, fails to work, and pretends to work. Routledge.

[52] Suraworachet, W., Seon, J., & Cukurova, M. (2024). Predicting challenge moments from students' discourse: A comparison of GPT-4 to two traditional natural language processing approaches. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 473-485).

[53] Tuval-Mashiach, R. (2021). Is replication relevant for qualitative research? *Qualitative Psychology, 8*(3), 365–377. https://doi.org/10.1037/qup0000217

[54] Wang, D., & Shang, Y. (2014). A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)* (pp. 112-119). IEEE.

[55] Xu, Z., Olson, J., Pochinki, N., Zheng, Z., & Yu, R. (2024). Contexts Matter but How? Course-Level Correlates of Performance and Fairness Shift in Predictive Model Transfer. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 713-724).

[56] Yan, L., Martinez-Maldonado, R., Zhao, L., Deppeler, J., Corrigan, D., & Gasevic, D. (2022). How do teachers use open learning spaces? Mapping from teachers' socio-spatial data to spatial pedagogy. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 87-97).

[57] Yan, L., Zhao, L., Gasevic, D., & Martinez-Maldonado, R. (2022). Scalability, sustainability, and ethicality of multimodal learning analytics. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 13-23).

[58] Zambrano, A. F., & Baker, R. S. (2024). Long-Term Prediction from Topic-Level Knowledge and Engagement in Mathematics Learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 66-77).

[59] Zambrano, A. F., Zhang, J., & Baker, R. S. (2024). Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 349-359).

[60] Zhang, Q., & Rutherford, T. (2022). Grade 5 Students' Elective Replay After Experiencing Failures in Learning Fractions in an Educational Game: When Does Replay After Failures Benefit Learning? In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 98-106).

[61] Zhao, L., Swiecki, Z., Gasevic, D., Yan, L., Dix, S., Jaggard, H., … & Martinez-Maldonado, R. (2023). METS: multimodal learning analytics of embodied teamwork learning. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 186-196).