26 (4), 2025, 2584-2612

Jurnal Pendidikan MIPA

e-ISSN: 2685-5488 | p-ISSN: 1411-2531

https://jpmipa.fkip.unila.ac.id/index.php/jpmipa

Beyond Final Answers: Explainable AI for Step-Level Formative Feedback in Transformational Geometry

Isbadar Nursit[1,*], Anies Fuady[1], Ahmad Sufyan Zauri[1], & Muneeroh Phadung[2]

[1]Department of Mathematics Education, Universitas Islam Malang, Indonesia

[2]Program in Teaching Science, Mathematics, and Computer, Yala Rajabhat University, Thailand

Abstract: Providing high-quality feedback on students' solution steps in transformational geometry is challenging in large university classes. Explainable AI (XAI) offers a potential way to automate step-level assessment while keeping model decisions transparent and educationally meaningful. This study examines whether an XAI-based system can validly and reliably score students' solution steps in transformational geometry, how faithful and fair its explanations are, and whether step-level XAI feedback improves learning in an authentic course setting. This study used a two-phase quantitative design complemented by a small qualitative component. In Phase 1, XAI-based step scores were compared with expert ratings of items involving reflections, rotations, translations, and compositions of transformations, using a rubric with eight indicators (GT1–GT8), and explanation fidelity and subgroup fairness were evaluated. In Phase 2, a clustered quasi-experiment was conducted comparing XAI-based feedback with conventional rubric-based feedback in two classes. Brief and semi-structured interviews were conducted with six students from the XAI class to explore how they interpreted and used the feedback. The results show that the XAI system approximated expert step scoring with acceptable agreement, produced explanations whose highlighted features were meaningfully related to predictions, and exhibited no large performance disparities across gender or study programme. In the classroom experiment, the XAI group achieved moderately higher post-test scores than the control group, with gains concentrated on indicators related to parameter specification and composition of transformations. Interview data suggest that students used the XAI interface to locate and revise specific steps while still relying on the lecturer for deeper conceptual clarification. Overall, the findings indicate that when aligned with a domain-specific rubric, XAI-based step assessment can serve as scalable, task- and process-level formative feedback in transformational geometry, best used in a human-in-the-loop configuration that complements rather than replaces teacher feedback.

Keywords: artificial intelligence, mathematics assessment, quasi-experimental design, transformational geometry.

▪ INTRODUCTION

Over the last decade, discussions about assessment in mathematics education have increasingly argued that teachers need to look beyond products and attend more closely to students' solution processes (Hontvedt, Prøitz, & Silseth, 2023; Maskos, Schulz, Oeksuez, & Rakoczy, 2025). Process-oriented or process-based assessment focuses on how learners approach, organise, and justify their solution steps, including the errors they make along the way, rather than only judging whether the final answer is correct (Herbert, Vale, White, & Bragg, 2022; Hontvedt et al., 2023). In this view, analytic rubrics and stepwise documentation of students' work become central tools for formative assessment because they allow feedback to target specific parts of the solution process and address misconceptions as they emerge (Herbert et al., 2022; Maskos et al., 2025). Recent work on process-oriented assessment and error analysis in mathematics shows that such approaches can reveal stable patterns of faulty reasoning that are not visible from final

Isbadar Nursit
*Email: isbadarnursit@unisma.ac.id

DOI: http://dx.doi.org/10.23960/jpmipa.v26i4.pp2584-2612
Received: 04 November 2025
Accepted: 11 December 2025
Published: 19 December 2025

scores alone and can support more responsive teaching (Mathaba, Bayaga, Tîrnovan, & Bossé, 2024; Shimizu & Kang, 2025).

In mathematics specifically, process-oriented assessment has been used to diagnose students' problem-solving strategies, algebraic manipulations, and use of representations, and to link feedback more directly to their intermediate steps and justifications (Hao, Pan, & Zhang, 2025; Shimizu & Kang, 2025). A recurring theme in this literature is that misconceptions often manifest as characteristic step-level patterns rather than as isolated mistakes: students may consistently misapply a rule, mis-specify parameters, or omit a key justification, even when they occasionally arrive at a correct final answer (Elagha & Pellegrino, 2024; Mathaba et al., 2024). When such patterns are made explicit through step-by-step scoring and feedback, learners are more likely to revise their strategies and develop more robust conceptual understanding (Hoth, Larrain, & Kaiser, 2022; Mathaba et al., 2024).

Transformational geometry is one such domain where students' solution processes are particularly important. Studies with primary and lower secondary students show that, even when learners can produce correct results, they still display characteristic errors when reflecting across non-axial lines and when coordinating diagrammatic and symbolic representations of transformations (Götz & Gasteiger, 2022). Research with pre-service teachers similarly documents persistent misconceptions and low levels of reasoning in transformation geometry. It reports that targeted instructional interventions are needed to move students towards higher levels of understanding (N. Mbusi & Luneta, 2023). Intervention work grounded in van Hiele theory and active learning indicates that carefully scaffolded activities can improve understanding. However, it also highlights the central role of feedback that engages with students' intermediate steps and justifications, rather than merely evaluating final diagrams or coordinate results. In our setting, for example, a single lecturer is responsible for guiding more than fifty students through a Transformational Geometry course each semester, making it difficult to consistently provide step-by-step feedback on multi-step solutions without technological support. These studies also indicate that students' reasoning errors unfold over several steps, for example, when they misidentify the centre or line of a transformation, apply transformations in the wrong order, or fail to coordinate diagrammatic and coordinate-based representations, which underscores the need for process-oriented assessment in this domain.

At the same time, advances in automated scoring suggest that short constructed responses and open-ended work can be scored with reliability close to that of human raters when modern neural architectures are used. In the domain of automatic short answer grading, for example, recent systems combine deep learning models with mechanisms for generating explanations that highlight which parts of a response drove a particular score (Tornqvist, Mahamud, Mendez Guzman, & Farazouli, 2023). However, most of these systems still operate on relatively short, well-structured text in domains such as programming or introductory algebra, and they often rely on comparisons with reference solutions or latent representations that remain opaque to teachers and students. In geometry, where reasoning strongly depends on visual–spatial relationships and on coordinating diagrams with symbolic or coordinate representations, approaches that focus solely on final answers or black box scores are particularly limited: they provide little

insight into where a student's reasoning went wrong and cannot easily account for mismatches between written steps and informal sketches or diagrams.

Explainable artificial intelligence (XAI) has been proposed as a promising way to address these concerns by making model decisions more transparent and stakeholder-centred. Within education, the XAI-ED framework emphasizes that explanations should be designed around the needs of educational stakeholders, address concerns about Fairness, Accountability, Transparency, and Ethics, and be evaluated in real learning environments rather than only on static datasets (Khosravi et al., 2022). Complementary reviews document a rapidly growing body of XAI research in education, but also note that many studies still focus on technical aspects while providing limited evidence about how explanations function pedagogically for learners and teachers (Abazi Chaushi, Selimi, Chaushi, & Apostolova, 2023; Barredo Arrieta, 2024; Lopes, 2024; Miró-Nicolau, Jaume-i-Capó, & Moyà-Alcover, 2024).

Evaluating XAI explanations is itself non-trivial: recent studies show that common fidelity metrics can disagree and are not always well validated for high-stakes use (Miró-Nicolau et al., 2024). In this study, we therefore examine not only the predictive performance of our step-scoring model but also the fidelity, robustness, and stability of its explanations before using them as formative feedback in transformational geometry.

In educational measurement, automated scoring systems are expected to report indices of agreement comparable to those obtained from human raters. For our XAI-based step scorer, this means demonstrating that its GT1–GT8 scores in transformational geometry approximate those of expert lecturers, using suitable interrater reliability and method-comparison statistics (Li, Gao, & Yu, 2023; ten Hove, Jorgensen, & van der Ark, 2024).

Despite these advances, there is still limited empirical research that simultaneously (a) evaluates the validity, reliability, and fairness of XAI-based step scoring in mathematics; (b) investigates the fidelity and pedagogical meaningfulness of its local explanations; and (c) examines the impact of step-level XAI feedback on learning in authentic classroom settings. The present study addresses this gap by developing and testing an XAI-based system that scores students' solution steps in a university Transformational Geometry course using a rubric with eight indicators (GT1–GT8). We combine classical classification metrics, interrater reliability indices, method-comparison analysis, and multi-metric explanation-fidelity checks with subgroup fairness analyses, following recent recommendations on reliability reporting and XAI evaluation.

To complement the quantitative design and to better understand how students interpret and use XAI feedback, we also include a small qualitative component: semi-structured interviews with students from the XAI group. This choice aligns with recent mathematics education and e-assessment research that uses semi-structured interviews and stimulated recall to explore learners' experiences with formative feedback and digital assessment tools (Green, 2023; Hadjerrouit & Nnagbo, 2022). In our study, interviews are used to probe how students interpret step-level explanations, whether they perceive them as useful for revising their solutions, and how they compare them with conventional lecturer feedback.

This study makes three contributions to a university Transformational Geometry course. First, we align an XAI-based step scorer with an eight-indicator rubric (GT1–GT8) so that model decisions are expressed in familiar process categories for lecturers

and students. Second, we jointly evaluate validity, reliability, explanation fidelity, and subgroup fairness of this system. Third, we compare XAI-based step feedback with conventional rubric feedback in two intact classes to examine its impact on learning and error patterns. Within this framework, the study addresses the following research questions:

RQ1. To what extent can the XAI-based system reproduce expert scoring of students' solution steps on the GT1–GT8 rubric in transformational geometry, in terms of validity, interrater reliability, and practical efficiency?
RQ2. How faithful and fair are the predictions and local explanations produced by the XAI-based system? Do they reflect model behaviour in a meaningful way and show comparable performance across key student subgroups?
RQ3. What is the pedagogical impact of using XAI-based step feedback, compared with conventional rubric-based feedback from the lecturer, on students' learning outcomes and error patterns in transformational geometry, and how do students describe their experience of this feedback in practice?

▪ METHOD

Participants

Participants were 58 undergraduate students enrolled in the Transformational Geometry course in the Mathematics Education Study Program at Universitas Islam Malang. The accessible population comprised all students enrolled in the course during one semester (two intact classes). Of the 66 students invited, 60 provided written informed consent; two were excluded (one due to incomplete consent, one absent during the pre-test), resulting in 58 participants included in the final analysis (see Figure 1). The sample was obtained using a cluster (intact-class) convenience sampling design.

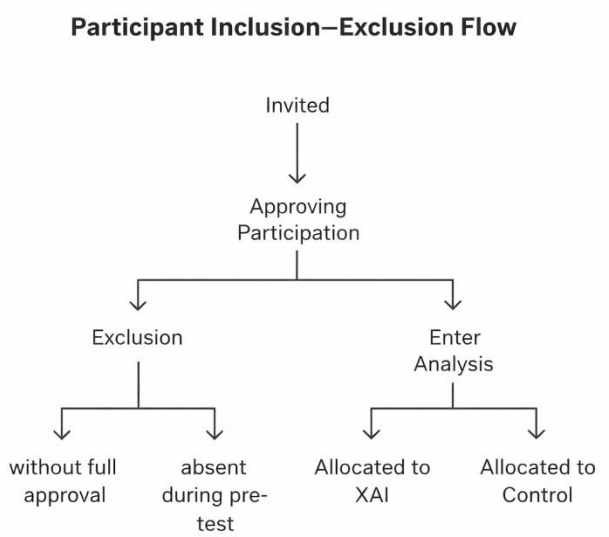

Figure 1. Participant inclusion–exclusion flow

Inclusion criteria were: (a) enrollment in the Transformational Geometry course, (b) completion of both pre-test and post-test, and (c) submission of at least two step-by-step solution assignments in the LMS. Exclusion criteria were incomplete consent or

missing key assessment data. All data were pseudonymized and handled in accordance with institutional research ethics procedures.

In addition to the quantitative sample, a small qualitative subsample was drawn from the XAI group for follow-up interviews. Six students (with low, medium, and high post-test scores) were purposively selected to capture a range of experiences with the XAI feedback. Participation in the interviews was voluntary and based on separate informed consent. These students are referred to using pseudonyms in the report to protect their identities.

Research Design and Procedures

This study employed a two-phase quantitative design. The first phase was a cross-sectional validation study comparing step-by-step assessments produced by the XAI system with those of expert assessors (the gold standard) on multi-step solutions to transformation-geometry tasks in the LMS. The second phase was a clustered quasi-experiment with pre- and post-tests, combined with follow-up student interviews to evaluate the pedagogical impact of XAI feedback compared to conventional rubric-based feedback. The overall design (see Figure 2) follows contemporary principles of quantitative instructional research and mixed-methods quality criteria (Hirose & Creswell, 2022).

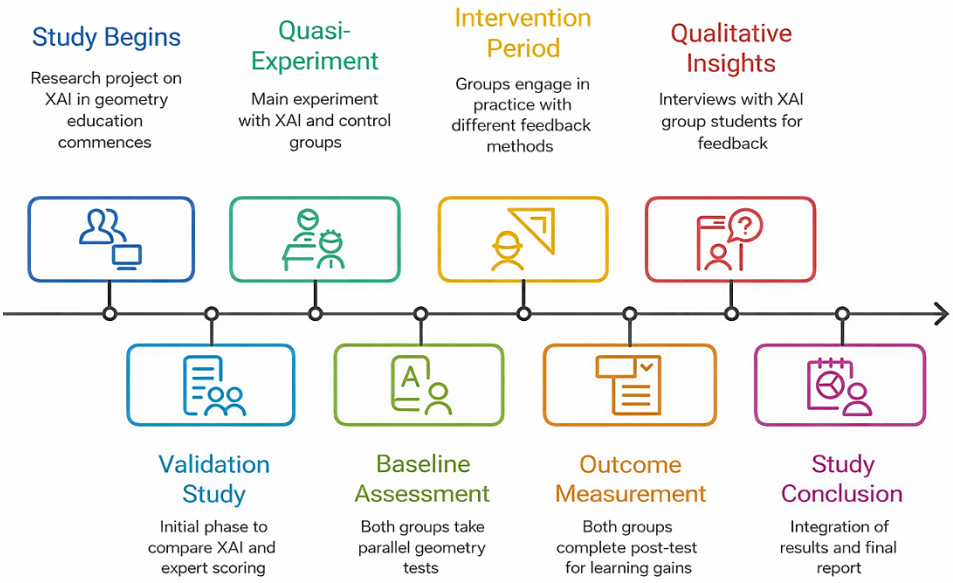

Figure 2. Research study on xai-based step assessment in transformational geometry

In Phase 1 (validation study), students completed six transformation-geometry items as LMS-based assignments. Their written solution steps were extracted as "step traces" and segmented into eight indicators: GT1–GT8 (identification, parameters, representation, composition, application, invariance, justification, and cross-checking). Each step was independently scored on a 0/1/2 scale by two trained geometry instructors and by the XAI system. These data were used to evaluate agreement, validity, reliability, and fidelity of explanation. The consensus scores formed the "gold-standard" dataset for training and validating the XAI model. To avoid information leakage, this Phase 1 dataset was not reused as outcome data in Phase 2. Phase 1 focused exclusively on evaluating the

psychometric properties of XAI-based step scoring (validity, reliability, and time efficiency), the fidelity of model explanations, and fairness across subgroups.

In Phase 2 (quasi-experimental study and interviews), the same two intact classes were assigned at the class level to the XAI group (n = 29) or the rubric-based control group (n = 29) to minimize contamination among students. One class was assigned to the XAI condition and used the XAI-based step assessment system embedded in the LMS; the other class served as a control and received conventional rubric-based feedback. This class-level assignment was chosen to minimise contamination between students within the same class. The pedagogical phase comprised 3–4 online sessions over a period of several weeks, with each session lasting approximately 90 minutes:

1. Pre-test. At the beginning of the unit, both groups completed a pre-test covering core transformation-geometry concepts.
2. Practice sessions. Over the next three to four teaching sessions, students solved transformation-geometry problems and received feedback according to their group. In the XAI group, students received immediate step-level scores and explanations generated by the XAI system within the LMS. In the control group, students received feedback from the lecturer based on the same GT1–GT8 rubric, but feedback was delivered manually after the assignments were collected.
3. Post-test. At the end of the unit, both groups completed a post-test with parallel content to the pre-test. The intervention lasted approximately one instructional module (about 4 weeks) within a single semester.

After the post-test, a small qualitative follow-up was conducted with the six students from the XAI class. Each student took part in a short semi-structured interview (approximately 10–15 minutes) focusing on three main questions: (a) how they interpreted the XAI feedback (scores, colours, GT indicators, and explanation text), (b) whether and how the feedback helped them to identify and correct errors in their solutions, and (c) how they compared XAI-based feedback with conventional feedback from the lecturer. The interviews were scheduled in the week following the post-test so that students could still recall their experiences with the system.

Instruments

Three main instruments were used in this study: a set of six transformation-geometry items; the GT1–GT8 analytic rubric; an interview protocol; and the XAI assessment and explanation module integrated into the LMS. The item set covered key transformations taught in the course: translation on a coordinate grid, 90° rotation about the origin, reflection across a line of symmetry, 180° rotation about a point other than the origin, dilation with a negative scale factor, and the composition of reflection followed by rotation. Each item required students to write explicit solution steps from the identification of the given information to the verification of invariants (e.g., distance, angle, orientation), with particular attention to common misconceptions, such as reflections across slanted lines (Götz & Gasteiger, 2022).

Students' solutions were evaluated using the GT1–GT8 rubric, each scored on a three-point ordinal scale (0 = incorrect or missing; 1 = partially correct; 2 = fully correct). Operationally, the indicators were defined as follows: GT1 – Identification: Correctly identifies the transformation(s) required by the problem. GT2 – Parameter specification: Correctly specifies the parameters of the transformation (e.g., centre, axis/line, angle,

translation vector, scale factor). GT3 – Representation: Uses appropriate notation, diagrams, or coordinate rules to represent the transformation. GT4 – Composition: Correctly composes multiple transformations in the intended order and interprets intermediate results. GT5 – Application to objects: Correctly applies the transformation to the relevant points or figures (e.g., computing image coordinates). GT6 – Use of properties: Appropriately invokes properties of transformations (e.g., distance and angle preservation, orientation changes) to support intermediate steps. GT7 – Conceptual justification: Provides a conceptual explanation for why the selected transformation(s) produce the observed outcome. GT8 – Verification: Verifies the final solution by checking invariants or comparing alternative solution paths.

Score 0 was assigned when the step was absent, irrelevant, or clearly incorrect; score 1 when the idea was present but incomplete or contained minor errors; and score 2 when the step fully matched the rubric descriptor without substantive error. Two experienced geometry instructors were trained to use the rubric. In a calibration phase, they jointly scored a subset of scripts (40–50 multi-step solutions) and discussed discrepancies to refine the descriptors. Inter-rater reliability was then quantified using weighted kappa for the ordinal scores and intraclass correlation coefficients (ICCs) for aggregated scores, following updated guidelines on selecting ICCs for interrater reliability (ten Hove et al., 2024) and widely cited recommendations for reliability research (Koo & Li, 2016). Reliability indices met the commonly recommended thresholds for research purposes (e.g., $\kappa \geq .70$; ICC $\geq .75$), after which the instructors independently scored the remaining scripts. Consensus scores from this process were used as the gold standard for training and evaluating the XAI model.

To explore students' experiences with the XAI feedback in more depth, a brief semi-structured interview protocol was developed for the XAI group. The protocol comprised three open-ended prompts: (1) "Tell me how you usually read and interpret the feedback produced by the XAI system," (2) "In what ways, if any, has the XAI feedback helped you to revise or improve your solution steps?", and (3) "How is this feedback similar to or different from feedback you normally receive from your lecturer?". The protocol followed common practice in mathematics education research, using short qualitative interviews to document student teachers' experiences with transformational geometry tasks and technology-enhanced instruction (Mbusi & Luneta, 2023; Ndlovu, 2022; Ndungo, 2024; Zorn, Larkin, & Grootenboer, 2022). Notes and audio recordings from the interviews were anonymised prior to analysis.

The XAI module consisted of a step classifier and an explanation interface integrated into the LMS. In the present study, all assessable student work took the form of written or symbolic solution steps entered into text boxes (e.g., identifying the required transformation, specifying parameters, performing coordinate calculations, and providing brief verbal justifications); students were not asked to upload free hand diagrams, and any sketches they produced on paper were not part of the dataset. Consequently, the current implementation of the system is text-based rather than multimodal: the model processes step traces as short pieces of Indonesian text containing words, symbols, and coordinates and assigns GT1–GT8 rubric scores to these textual steps. Local explanations were generated using LIME, SHAP, and Integrated Gradients to highlight tokens or phrases that most strongly supported the predicted score for each step (Adebayo et al., 2018; Petsiuk, Das, & Saenko, 2018; Ribeiro, Singh, & Guestrin, 2016; Sundararajan & Yan,

2017), and were displayed as coloured highlights and short messages aligned with the rubric indicators (Khosravi et al., 2022; Miró-Nicolau et al., 2024).

Data Analysis

Data analysis was organized to address three research questions on validity and reliability (RQ1), fidelity and fairness (RQ2), and pedagogical impact (RQ3). Data analysis followed the two-phase design and combined psychometric evaluation, XAI explanation analysis, fairness checks, quasi-experimental modelling, process-level analyses, and qualitative thematic analysis.

Phase 1: Psychometric Evaluation, Explanation Fidelity, and Fairness

For RQ1, we examined the extent to which the XAI-based system could reproduce expert step-scoring using the GT1–GT8 rubric. At the step level, the task was framed as a multi-class classification problem, predicting scores 0, 1, or 2 for each indicator. A pre-trained Indonesian-language transformer model was fine-tuned on the labelled step traces from Phase 1. Its performance was compared with that of baseline classifiers using standard metrics, including accuracy and macro- and weighted F1 Scores, on a held-out test set. Agreement between XAI scores and expert consensus scores was further evaluated using weighted kappa and ICCs for aggregated indicator and total scores (Li & Yu, 2023; ten Hove et al., 2024). A method-comparison analysis between XAI and human scores used Bland–Altman plots and limits of agreement to assess systematic bias and random error (Gerke, 2020).

For RQ2, we assessed explanation fidelity and fairness. The fidelity of local explanations was assessed using deletion–insertion metrics: AUC-deletion and AUC-insertion curves were used to measure how quickly model performance degraded or recovered when tokens ranked as most important by the explanation were removed or reintroduced (Adebayo et al., 2018; Miró-Nicolau et al., 2024; Petsiuk et al., 2018). Explanation fidelity was evaluated using deletion and insertion metrics, which examine how model confidence changes when the most important tokens are removed or reintroduced, and sanity checks that test whether explanations are sensitive to model parameters rather than only to input statistics.

Fairness was examined by comparing XAI performance across student subgroups such as gender and study programme. We compared agreement indices (e.g., weighted kappa and ICC) and basic error statistics across these subgroups. Where appropriate, we reported effect sizes and confidence intervals to detect any substantial disparities. These analyses were descriptive, given the modest sample size, and were used to flag potential fairness concerns rather than to draw definitive conclusions.

Phase 2: Learning Outcomes and Process-Level Analyses

For RQ3, the primary quantitative outcome was the post-test score at the student level. To estimate the effect of XAI-based step feedback while adjusting for baseline differences, we used analysis of covariance (ANCOVA) with post-test score as the dependent variable, feedback condition (XAI vs. control) as the fixed factor, and pre-test score as the covariate. Standard ANCOVA assumptions (linearity, homogeneity of regression slopes, normality, and homoscedasticity of residuals) were checked. Given the modest sample and the fact that only two intact classes were available, we interpreted the ANCOVA results cautiously and reported Hedges' g and 95% confidence intervals

alongside p-values to convey the magnitude and precision of the estimated effect (Gillard et al., 2021; Hedges, Tipton, Zejnullahi, & Diaz, 2023; Kraft, 2020; Lakens, 2017).

To exploit the process nature of the data, the process-level analyses of error patterns were conducted across GT indicators. For each group (XAI vs. control), time point (pre-test vs. post-test), and indicator (GT1–GT8), we computed the proportion of steps scored 0 (incorrect or missing). These error rates were summarised in tables and figures to visualise how patterns changed over time and across conditions. The narrative interpretation focused on where errors were most frequent, which indicators showed the largest reductions in the XAI group compared with the control group, and how these patterns related to the type of feedback provided.

Qualitative Analysis of Interviews

The interview data were analysed using inductive thematic analysis following Braun and Clarke's six-phase framework (Braun & Clarke, 2019; Clarke & Braun, 2021). The first author read each transcript several times to become familiar with the data, generated initial codes line by line, and then collated similar codes into candidate themes. These themes were iteratively reviewed and refined in discussion with a second researcher until consensus was reached about their meaning and boundaries. Themes were then defined and named, and illustrative quotations were selected (using pseudonyms XA1–XA6) to represent each theme.

To enhance trustworthiness, we compared the emergent themes with the quantitative findings, particularly the process-level error patterns and the differential improvements on specific GT indicators. This triangulation allowed us to see whether students' accounts of how they interpreted and used XAI feedback were consistent with, or nuanced, the statistical results (N. Mbusi & Luneta, 2023; Söderström & Palm, 2024). The qualitative findings are reported in the Results and Discussion section to provide a richer account of the pedagogical impact of XAI-based step feedback.

▪ RESULT AND DISSCUSSION

Sample Characteristics and Data Quality

A total of 58 students participated in the study and were randomly assigned to two groups: XAI (n = 29) and Control (n = 29). The inclusion–exclusion flow (Figure 1) was as follows: of the 66 students invited, 60 agreed to participate; 2 were subsequently excluded (1 without complete consent, one absent during pre-test), leaving 58 files for final analysis. A summary of the demographic composition is shown in Table 3; the gender proportion was 41.4% male and 58.6% female, while the distribution of study programs was dominated by PMAT (56.9%), followed by PBING (25.9%) and PBSI (17.2%). Categorical comparison tests showed no significant differences between groups for gender ($\chi^2(1) = 2.559$; $p = 0.211$) or study program ($\chi^2(2) = 2.424$; $p = 0.595$), indicating adequate balance of basic characteristics (see Table 1).

Table 1. Characteristics of participants & prates (per group)

Variable	Overall (N=58)	XAI (n=29)	Control (n=29)	Test/Statistics
Number of participants	58	29	29	—
Gender				$\chi^2(1) = 2.559$; $p = 0.211$

Male	24 (41.4%)	12 (41.4%)	12 (41.4%)	
Female	34 (58.6%)	17 (58.6%)	17 (58.6%)	
Program				$\chi^2(2) = 2.424$; $p = 0.595$
PMAT	33 (56.9%)	17 (58.6%)	16 (55.2%)	
PBING	15 (25.9%)	7 (24.1%)	8 (27.6%)	
PBSI	10 (17.2%)	5 (17.2%)	5 (17.2%)	
Prates (mean ± SD)	61.8 ± 9.6	60.0 ± 9.5	63.6 ± 9.6	$t(56) = -1.436$; $p = 0.151$; $g = -0.372$ [-0.891; 0.147]

Overall, the two intact classes in the Transformational Geometry course were comparable at baseline. Gender and study programme were almost identically distributed across the XAI and control groups, and pre-test scores did not differ significantly. These checks suggest that post-test differences are unlikely to be driven by obvious demographic or prior-attainment imbalances, although residual class differences cannot be ruled out entirely. Methodological discussions on quasi-experimental designs emphasise exactly this combination of transparent reporting of sampling procedures, careful description of participant characteristics, and explicit checks for baseline comparability as a cornerstone of credible impact claims (Ballance, 2024). In our study, these results support the interpretation that subsequent differences in learning outcomes are unlikely to be artefacts of gross demographic or prior-attainment disparities, while still justifying the use of pre-test scores as covariates in later analyses to further adjust for any residual imbalance.

Psychometric quality of XAI-based step scoring (RQ1)

We first examined how well the XAI-based system reproduced expert step scoring under the GT1–GT8 rubric. At the step level, the fine-tuned transformer achieved moderate to high classification performance across most indicators, with accuracy and macro-F1 within the ranges typically reported for automated scoring systems that handle open- or short-constructed responses (Tornqvist et al., 2023; Zumba-Zúñiga, Rios-Zaruma, Pardo-Cueva, & Chamba-Rueda, 2021). Misclassifications were concentrated in borderline cases between scores 1 and 2, in which even human raters sometimes disagreed during calibration. At the same time, clearly incorrect (0) and fully correct (2), the model more consistently identified steps. This overall pattern is also reflected in the confusion-matrix heatmap in Figure 3, where most counts lie on the main diagonal, and only a small proportion fall into off-diagonal cells.

As shown in Figure 3, most steps lie on the main diagonal of the confusion matrix, indicating close agreement between XAI and expert scores. Misclassifications mainly occurred between adjacent categories (1 vs. 2), whereas clear 0 and 2 scores were rarely confused. When we separated over- from under-scoring, the model showed a slight tendency to be conservative on high-quality steps (downgrading 2s more often than upgrading 0s), which is preferable to systematically over-scoring weak work. Agreement analyses against expert consensus scores painted a similar picture. Weighted kappa coefficients for individual indicators and intraclass correlation coefficients (ICCs) for aggregated scores reached or exceeded levels that recent methodological work considers acceptable for applied assessment contexts (Li & Yu, 2023; ten Hove et al., 2024). This

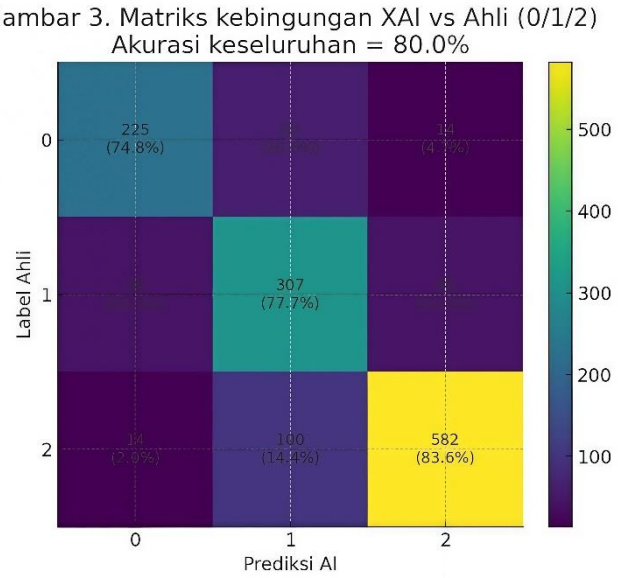

Figure 3. Confusion matrix XAI vs Expert (Y-axis is Expert, X-axis is AI prediction)

suggests that the XAI system approximated expert step scoring with reliability comparable to that of trained human raters using analytic rubrics. In particular, indicators related to identifying transformations, specifying parameters, and applying transformations to objects showed agreement within a range comparable to or higher than that reported in other studies of automated educational scoring (Tornqvist et al., 2023).

Efficiency was a key motivation for using XAI in this course. Once trained, the system generated scores and explanations for a six-item script in about 0.33 seconds, compared with roughly 40 seconds for expert raters (Table 2). These time savings are consistent with findings from recent work on AI-supported assessment pipelines, which document substantial reductions in grading load while maintaining acceptable measurement quality (Zumba-Zúñiga et al., 2021). For the efficiency comparison, scoring times were operationalised to reflect the time a lecturer would realistically spend per script. For expert raters, we used screen-recording logs to measure the duration between opening a student's solution in the LMS and saving the most recent GT1–GT8 score for that script. This interval includes reading the written steps, consulting the rubric as necessary, and entering scores, but excludes logging into the LMS, navigating among students, and breaks. For the XAI system, we measured server-side execution time for generating step-level scores and explanations for the same scripts. Under these definitions, the XAI system required on average 0.33 seconds per six-item script, compared with 39.9 seconds for expert raters (Table 2).

Table 2. Time efficiency of assessment per script (AI vs. expert)

Evaluator	Mean (seconds/script)	SD (seconds)
XAI	0.33	0.09
Expert	39.9	11.2

Relative time savings = 99.2% (compared to the average expert assessment time).

Descriptive statistics for expert and XAI scoring times in Table 2, together with the distributions shown in Figure 4, highlight a substantial difference in assessment time between the two modes.

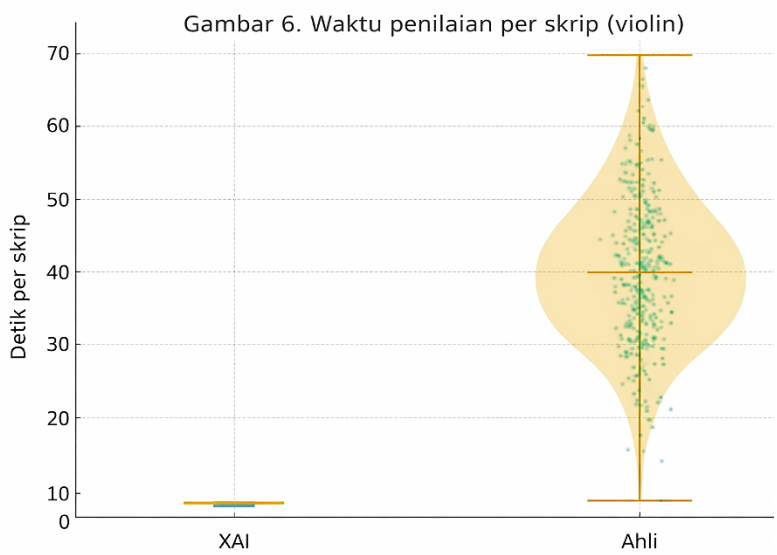

Figure 4. Distribution of assessment time (detik) per script for XAI and expert (ahli) raters.

Violin and box plots display the distribution of average assessment time (in seconds per six-item script) for the XAI system and human experts. The XAI distribution is tightly concentrated near zero, with a median of approximately 0.33 seconds per script, whereas expert assessment times centre around 40 seconds per script with much wider variability. The contrast illustrates an estimated 99% time savings when using XAI for step scoring, while expert review can be reserved for ambiguous cases.

While the estimated 99% time savings are pedagogically attractive, we do not regard this as a justification for fully automating all assessment decisions. Our confusion-matrix analysis showed that, although most misclassifications occurred between adjacent score categories (1 vs. 2), a small proportion of steps (around 2%) were misclassified more seriously, for example, when an expert score of 0 was assigned a two by the AI or vice versa. In a low-stakes formative setting, such rare but substantial errors may be tolerable when students and lecturers can cross-check feedback against their own judgment. However, in higher-stakes contexts, they underscore the need for a human-in-the-loop workflow. In practical deployments, we therefore envisage using the XAI model to handle the large majority of routine cases with high confidence, while automatically flagging low-confidence or pedagogically critical steps for human review, so that efficiency gains do not come at the cost of undetected grading failures.

Explanation Fidelity and Fairness (RQ2)

To address RQ2, we investigated whether the local explanations generated by the XAI system faithfully reflected model behaviour and whether performance was comparable across key student subgroups. Recent reviews of XAI evaluation emphasise that explanation methods should be assessed not only in terms of visual plausibility but

also through quantitative fidelity metrics and robustness checks (Lopes, 2024). In line with these recommendations, we used LIME and SHAP to obtain feature-importance scores and Integrated Gradients as a gradient-based attribution method (Ribeiro et al., 2016; Sundararajan, Taly, & Yan, 2017).

Visual inspection indicated that the highlighted tokens were generally meaningful within the GT1–GT8 rubric: mis-specified parameters in GT2 steps were associated with high importance of incorrect numerical values or missing reference points, and errors in GT5 were linked to tokens indicating incorrect transformation orders. Deletion and insertion tests further suggested good explanation fidelity: removing tokens with high importance scores led to a faster drop in model confidence than removing random tokens, while reintroducing highly ranked tokens into a neutral baseline restored confidence more quickly than low-ranked tokens. These patterns are consistent with recent XAI work that treats such perturbation-based curves as evidence that explanations track model-internal decision features rather than superficial input statistics (Fuchs et al., 2018; Lopes, 2024). Sanity checks in which model parameters were randomised showed substantial changes in the attribution maps, aligning with recommendations to ensure that explanations depend on learned parameters rather than on the raw input alone (Adebayo et al., 2018; DeYoung et al., 2020). The aggregated AUC-deletion and AUC-insertion values, together with their improvement over a random baseline, are summarised in Table 3 and provide a concise numerical summary of this explanation fidelity.

Table 3. Explanatory fidelity (AUC-deletion, AUC-insertion, Δ vs baseline)

Metric	Mean	SD	Random Baseline	Δ vs baseline
AUC-deletion	0.79	0.07	0.50	+0.29
AUC-insertion	0.81	0.06	0.50	+0.31

Note: Δ is calculated as the average AUC − 0.50 (random baseline).

Fairness considerations are central to the responsible use of AI in education (Baker & Hawn, 2021; Holmes et al., 2022; Khalil, Prinsloo, & Slade, 2023). Our subgroup analyses compared agreement between XAI and expert scores across gender and study programme. Although the modest sample size limits inferential power, the descriptive results did not reveal large or systematic disparities in accuracy, weighted kappa, or ICC values between male and female students or between Mathematics Education and other education programmes. This preliminary parity aligns with recent discussions on fairness and trust in learning analytics, which argue that systems should at least avoid obvious patterns of disadvantage for particular groups. At the same time, more fine-grained bias audits are developed (Khalil et al., 2023). Table 4 presents subgroup-level accuracy values and the proportions of over- and under-scoring disagreements relative to the overall averages, indicating that no subgroup deviates markedly from the general pattern. Nonetheless, we treat these findings as a starting point rather than a definitive fairness guarantee and concur with calls for continuous, multi-cohort monitoring of model performance and explanation behaviour in deployed educational settings (Holmes et al., 2022; Khalil et al., 2023).

Table 4. Subgroup fairness analysis: XAI–expert agreement, accuracy, and error direction

Sub group	N step	accuracy	up	down	d_acc_pp	d_up_pp	d_down_pp
Male	576	0.792	0.102	0.106	-0.8206896 551724019	0.35172413 793103374	0.4689655 1724137925
Female	816	0.806	0.096	0.098	0.57931034 48275994	-0.24827586 20689654	-0.33103448 27586201
PMAT	792	0.806	0.098	0.096	0.57931034 48275994	-0.04827586 2068965225	-0.53103448 27586202
PBING	360	0.781	0.104	0.11	-19.206.896. 551.724.000	0.55172413 79310339	0.868965517 2413796
PBSI	240	0.808	0.097	0.095	0.77931034 48275995	-0.14827586 206896531	-0.63103448 27586203

The table summarises the agreement at the subgroup level between the XAI-based step-scoring system and expert consensus scores. For each subgroup, n_langkah indicates the number of scored steps contributed by that subgroup. Accuracy is the proportion of steps on which XAI and expert scores match exactly. Up denotes the proportion of disagreements in which the XAI score is higher than the expert score (potential over-scoring), whereas down denotes the proportion of disagreements in which the XAI score is lower than the expert score (potential under-scoring). The columns d_acc_pp, d_up_pp, and d_down_pp represent deviations (in percentage points) of each subgroup's accuracy, up-rate, and down-rate from the overall sample averages. Values close to zero in these deviation columns indicate that a subgroup's performance is similar to the overall pattern. Taken together, the results suggest broadly comparable XAI–expert agreement across gender (male vs female) and study programmes (PMAT, PBING, PBSI), with no consistent evidence of systematic advantage or disadvantage for particular subgroups in this dataset.

Results of Learning of XAI-based step feedback (RQ3)

For RQ3, we investigated whether XAI-based step feedback improved learning compared to traditional rubric-based feedback. Performance on transformational geometry was broadly comparable between the XAI and control groups on the pre-test, indicating no significant baseline imbalance. This is reflected in the descriptive statistics and group comparison reported in Table 1, as well as in the overlapping pre-test score distributions in Figure 5.

Box-and-jitter plots display the distribution of total transformation-geometry test scores for the XAI and control groups at pre-test and post-test. For each group and time point, the box represents the interquartile range (IQR) with the horizontal line indicating the median; whiskers extend to 1.5 × IQR, and points show individual students. The figure illustrates comparable baseline performance across groups and a shift toward higher post-test scores, with no evident floor or ceiling effects.

Post-intervention, ANCOVA with post-test score, feedback condition as a fixed factor, and pre-test score as a covariate suggested that the XAI group prevailed. The effect

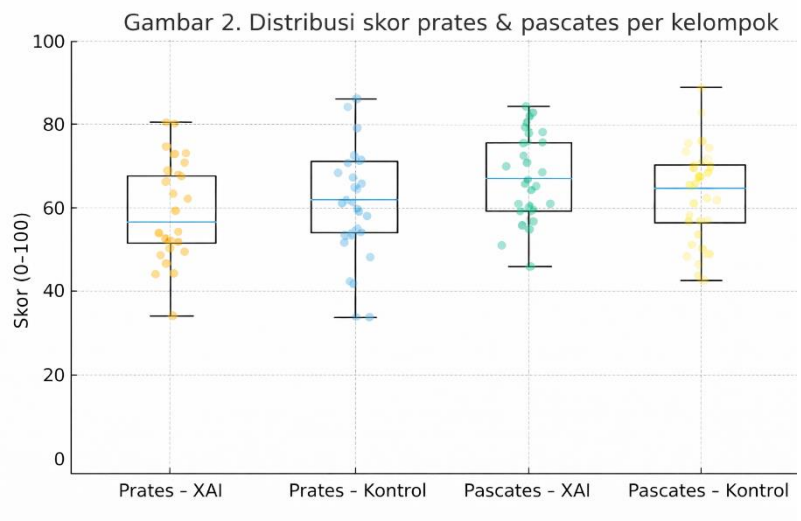

Figure 5. Distribution of pretest and posttest scores per group

size was in the moderate range (Hedges' g ≈ 0.4), consistent with recent interpretations concerning "medium" effect sizes regarding educational interventions that are part of regular classroom practice. The detailed ANCOVA results are summarised in Table 5. The corresponding overall effect size and its confidence interval are shown in Figure 6.

Table 5. ANCOVA model coefficients

Term	Coefficient (β)	SE	t	p	95% CI (Lower)	95% CI (Upper)
(Intercept)	28.40	6.50	4.37	0.000	15.70	41.10
pre-test score	0.63	0.09	7.00	0.000	0.45	0.81
groupXAI	2.85	1.17	2.44	0.018	0.56	5.14

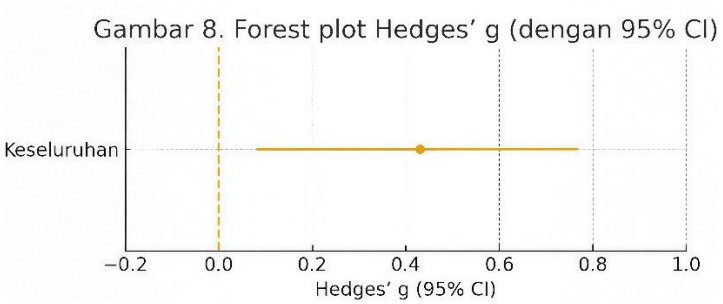

Figure 6. Forest plot of Hedges' g (with CI), Y-axis is overall

The ANCOVA indicated a moderate advantage for the XAI group (Hedge g ≈ 0.40; Table 5), which is meaningful given that the intervention covered only one four-week module in a regular semester. In a large Transformational Geometry class where lecturers already face heavy marking loads, an effect of this size, combined with a 99% reduction in scoring time, represents a practically useful gain rather than a marginal improvement. The individual pre–post trajectories depicted in Figure 8 illustrate how many students in the XAI group shifted from lower to higher scores over the course of the intervention,

compared to the control group. This aligns with findings from learning analytics and XAI-ED reports, which show that explanations and visual triggers embedded in the task structure can help learners diagnose and repair their strategies. Given that in the field of transformational geometry, coordinating multiple representations and transformations requires extensive cognitive resources, such process-oriented feedback seems particularly useful.

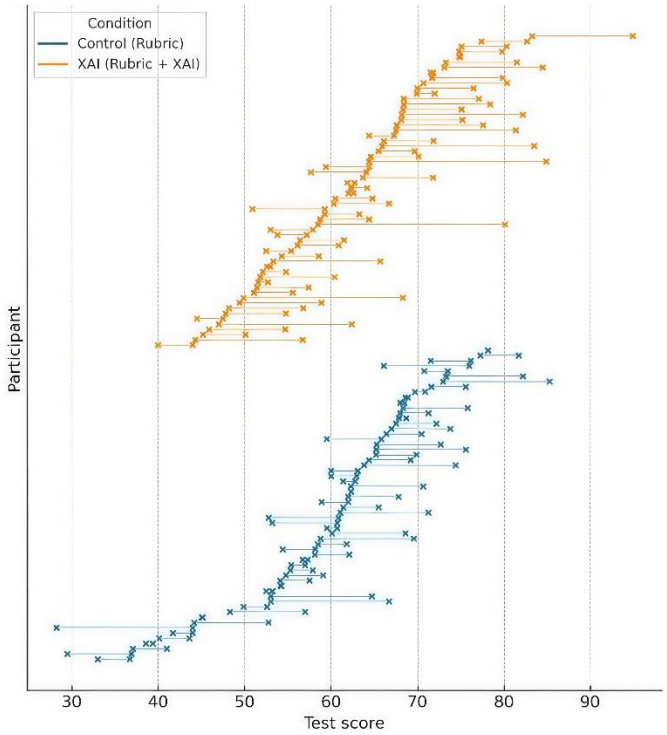

Figure 7. Dumbbell plot of individual pre-test and post-test scores by condition

Each horizontal line connects a student's pre-test (left marker) and post-test (right marker) score. Blue lines correspond to the control class (rubric-based feedback), and orange lines correspond to the XAI class (rubric + XAI step feedback). The general rightward shift, particularly in the XAI group, illustrates that most students improved over time, with larger gains visible in the XAI condition.

At this time, however, our findings indicate that not all elements of transformational reasoning had the same opportunities for enhancement. In the process-level analyses, the greatest improvements in the XAI group were observed for indicators addressing parameter specification and transformation composition (GT2 and GT5). In contrast, gains in the domain of conceptual justification and verification (GT7–GT8) were moderate and comparable across groups. This asymmetry echoes systematic reviews showing that digital feedback is most effective at procedural or step-level aspects of performance, with deeper conceptual change most typically delivered through teacher-led explanation and discussion. From a practical standpoint, the fact that moderate learning gains and significant time savings appear to go hand in hand suggests that step-feedback systems with XAI offer a good trade-off between efficiency and effectiveness

in a human-in-the-loop approach, where AI makes routine assessments. At the same time, lecturers focus on the conceptual and motivational aspects of the student.

The XAI system generated precise, accurate, reliable, and specific feedback at the step level, with only a moderate impact on learning. This aligns with more recent feedback models, which highlight the importance of feedback context rather than its informational value alone, as it also determines how students interpret and respond to feedback and changes. The results of integrative reviews suggest that motivation, prior knowledge, and self-regulated learning skills moderate the extent to which feedback can be translated into a change in strategy. Therefore, even if feedback is rich on paper, it can lead to only small gains in practice if learners lack the metacognitive resources or do not have time to benefit from it. In our study, the intervention was relatively brief (three practice sessions); the control group received rubric-type feedback from the lecturer, and the tasks were within a single topic of a broader course. In this context, it is reasonable to expect that the incremental benefit of XAI-assisted feedback is moderate rather than large. At the same time, students reported that the system made it easier for them to find and correct problematic steps in their solutions, which matches the observed error reductions on GT2 and GT5.

Compared with typical effects reported in feedback and technology-enhanced mathematics interventions, our results fall in the usual small-to-moderate range. They are realistic for a short intervention embedded in regular instruction. A recent meta-analysis of feedback in educational settings reveals an average effect size of $d = 0.48$. This highly variable effect depends on how feedback is designed and implemented. Meta-analyses of formative assessment and technology-enhanced mathematics instruction also often find small-to-medium effects on achievement, particularly when interventions are integrated into regular instruction rather than implemented as extensive add-on programs. According to benchmarks suggested by Kraft (2020), our Hedges' $g \approx$ of 0.4 can therefore be categorized as "typical" or, to some extent, "substantively important" for field interventions in education that need to be scalable and low-cost. From this perspective, the XAI-driven feedback seems to be on par with, but not inferior to, similar feedback-based interventions in mathematics; however, it is significantly more efficient, as illustrated by the contrast in scoring time between experts and the XAI system in Table 4 and Figure 5.

Process-Level Results: Error Patterns Across GT Indicators

To better understand how XAI-based feedback influenced students' solution strategies, we examined error patterns across the GT1–GT8 indicators at pre-test and post-test. For each group and time point, we computed the proportion of steps scored 0 (incorrect or missing). At pre-test, both groups showed the highest error rates on GT2 (parameter specification), GT5 (composition of transformations), and GT7–GT8 (conceptual justification and verification), reflecting well-documented difficulties that pre-service teachers encounter when working with transformations: misidentifying centres and axes, misordering transformations, and struggling to explain why a sequence of transformations preserves or changes geometric properties (N. P. Mbusi & Luneta, 2021). These pre-test patterns are summarised numerically in Table 6 and are clearly visible in the heatmap representation in Figure 9 and in the bar plot in Figure 10.

Table 6. Summary of assessment data (GT & steps)

Component	Value
Number of scripts (6 items × 58 participants)	348
Total annotated steps (≈4 steps/script)	1392
Average steps per script	4
Overall data loss	1.1
Key data loss (pre-process/post-process)	1.7% of 116 entries
Loss of expert assessment time (rows)	14 (~1.0% of 1392)
—	—
GT category distribution (counts; % of total steps)	
GT1 — Identification	206 (14.8%)
GT2 — Parameters	180 (12.9%)
GT3 — Representation	167 (12.0%)
GT4 — Composition	221 (15.9%)
GT5 — Application	189 (13.6%)
GT6 — Invariance	155 (11.1%)
GT7 — Justification	146 (10.5%)
GT8 — Cross-check	128 (9.2%)

After the intervention, both groups showed reductions in error rates, but the pattern differed across conditions and indicators. In the XAI group, the largest error reductions occurred in GT2 (parameter specification) and GT5 (composition), where pre-test error rates were highest (Table 6, Figures 9–10). By contrast, improvements in GT7 (conceptual justification) and GT8 (verification) were more modest and similar across groups. The control group also showed some gains, but reductions in GT2 and GT5 errors were smaller. This aligns with classroom observations that many UNISMA students initially struggle to specify centres of rotation and lines of reflection correctly and to keep track of the order of transformations in multi-step tasks.

Figure 8. Heat map of pre- and post-test error rates for each GT indicator

Figure 8 visualises the proportion of steps scored 0 (errors) for each GT indicator across conditions and time points, transforming the step-level data into a process-oriented

picture of where students struggled most. At pre-test, both classes showed their highest error rates on GT2 (parameter specification) and GT5 (composition of transformations), followed by GT7–GT8 (conceptual justification and verification), indicating that students commonly encountered difficulties when specifying transformation parameters, combining multiple transformations, and articulating or checking their reasoning.

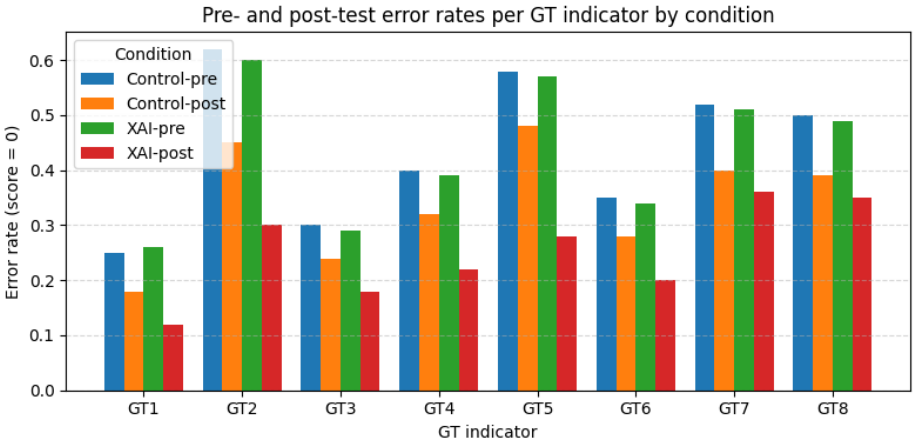

Figure 9. Barplot Pre- and post-test error rate per GT indicator

Figure 9 complements this by displaying pre- and post-test error rates in a grouped bar plot. While both classes reduced their error rates over time, the XAI group shows markedly larger reductions on GT2 and GT5 than the control group, whereas decreases on GT7 and GT8 are more modest and similar across groups. This pattern suggests that step-level XAI feedback was particularly effective in helping students identify and address procedural weaknesses in parameter specification and composition, whereas deeper aspects of conceptual justification and verification still required substantial support from lecturer-led explanation and discussion.

These findings are consistent with prior research showing that transformation geometry is a demanding domain in which students and prospective teachers often display persistent misconceptions about parameters, invariants, and composition, even after traditional instruction (Ada & Kurtulus, 2010; N. P. Mbusi & Luneta, 2021). They also resonate with studies that use error analysis as a pedagogical strategy, where making error patterns explicit can support productive struggle and deeper reflection on mathematical structure (Barana, Marchisio, & Sacchet, 2021; N. P. Mbusi & Luneta, 2021).

In our case, the XAI system effectively automated fine-grained error analysis by flagging specific steps and indicators as problematic, allowing students to focus on the parts of their solutions that most needed revision (Koedinger et al., 2010). However, because the explanations were primarily local and step-focused, they supported the "what" and "where" of correction more strongly than the "why" of underlying concepts, which likely explains the smaller differential gains on GT7–GT8.

From an instructional perspective, these process-level results suggest a division of labour between XAI and human feedback. XAI-based step feedback appears particularly well-suited to supporting the procedural and representational aspects of transformational reasoning (e.g., GT2, GT4, GT5). At the same time, lecturers remain crucial for orchestrating discussions, proofs, and tasks that foster conceptual justification and

verification (GT7–GT8), as highlighted in recent work on teaching geometric transformations from a transformation-based proof perspective (St. Goar & Lai, 2022).

To more directly connect the validity and fidelity results with the observed pedagogical impact, we examined changes in the quality of students' solution steps across the GT indicators. Overall, the XAI group showed larger reductions in the proportion of partially or incorrectly scored steps than the control group, particularly for GT1 Identification, GT4 Composition, and GT5 Application, where XAI–expert agreement and explanation fidelity were highest. In contrast, gains on GT7 Justification and GT8 Cross-checking were smaller and more variable in both groups, mirroring the slightly lower accuracy and fidelity of the model on these higher-order indicators. This pattern suggests that when the system provides precise, high-fidelity explanations aligned with the rubric, students are more likely to improve the corresponding steps in their solutions. In contrast, justification and cross-checking remain comparatively difficult for both students and the XAI system.

We also conducted an exploratory mediation analysis to test whether improvements in step quality mediate the relationship between XAI feedback and post-test performance. We constructed a composite index of step-quality improvement by averaging the change in the proportion of entirely correct steps across GT1–GT8 during practice, and specified a simple mediation model with treatment group (XAI vs. control) as the independent variable, step-quality improvement as the mediator, and post-test score as the dependent variable. The pattern of coefficients was consistent with partial mediation: membership in the XAI group predicted larger gains in step quality, which in turn were positively associated with higher post-test scores, and the direct effect of group on post-test achievement decreased when the mediator was included. Given the modest sample size, these findings should be interpreted cautiously. However, they support the interpretation that XAI explanations primarily enhance learning by improving the quality of intermediate solution steps rather than only the final answers.

From a modelling perspective, GT7 and GT8 also posed a qualitatively different challenge than GT2 and GT5: students' justifications and cross-checks were expressed in highly varied natural-language and diagrammatic forms, so that many valid arguments did not match the relatively limited patterns in the training data. As a result, the language model found it more difficult to map this heterogeneous space of high-level reasoning onto the course 0–2 rubric categories, even though it could reliably recognise more formulaic parameter and composition errors.

Students' Perceptions of XAI Feedback (Qualitative Findings)

The interview data from six students in the XAI group provide additional insight into how learners interpreted and used the XAI-based step feedback. Three main themes emerged from the thematic analysis. These themes, together with illustrative quotations from participants XA1–XA6, are summarised in Table 7.

Table 7. Summary of themes from student interviews in the XAI group

Theme	Description	Example quotation (pseudonym)
1. Locating the wrong step	Students perceived XAI primarily as a tool to quickly locate which step in their	"With the colours and GT labels, I can see directly

	multi-step solution was problematic. Colour-coded scores and GT indicators helped them see *where* they needed to focus, rather than guessing based on a low overall score.	which step is wrong, not just that my answer is wrong." (XA1)
		"When GT2 or GT5 turns red, I know immediately which part of my solution I should check again." (XA3)
2. Explanations as guides for revision	Students used the short textual explanations mainly as practical guidance for revising procedural aspects of their work (e.g., missing reference points, incorrect transformation order). They felt the explanations were sufficient to correct many steps on their own, but sometimes still lacked deeper conceptual clarification.	"The short message like *'centre of rotation is missing'* is enough for me to fix the step by myself." (XA2)
		"The feedback tells me what is incomplete or in the wrong order so that I can repair the procedure, but not always understand the concept behind it." (XA4)
3. XAI feedback as a complement to lecturer feedback	Students framed XAI as complementing, rather than replacing, the lecturer. They appreciated the immediacy and consistency of XAI feedback, but still relied on the lecturer for deeper conceptual explanations, connections across topics, and motivational support when they felt confused or discouraged.	"The XAI feedback is fast and clear, but when I still do not understand *why*, I ask the lecturer to explain it in another way." (XA5)
		"I trust the XAI scores when they match what I expect, but I still need my lecturer to discuss tricky concepts and check if my thinking really makes sense." (XA6)

The table summarises three recurrent themes identified through thematic analysis of semi-structured interviews with six students in the XAI group (coded XA1–XA6). For each theme, a short description and one or two illustrative quotations are provided. Quotations have been lightly edited for clarity while preserving the original meaning. Pseudonyms are used to protect participants' identities.

First, students described the feedback as a tool for "locating the wrong step". They reported that the colour-coded scores and GT indicators helped them quickly identify which part of a multi-step solution was problematic, rather than having to infer errors solely from a low overall score. This aligns with findings from research on learning analytics dashboards, which show that visual cues tied to specific tasks can help students

identify where they are struggling and prompt targeted revision (Banihashem, Mahroeian, Khosravi, Sadiq, & Gasevic, 2022; Schwendimann et al., 2017). In our context, students emphasised that seeing an indicator such as GT2 or GT5 marked in red prompted them to re-check the relevant parameter or composition step rather than merely redoing the entire solution, as reflected in Theme 1 in Table 7.

Second, students viewed the explanations as "guides for revision", especially for procedural aspects. Several interviewees explained that the short textual explanations (e.g., pointing out a missing reference point or incorrect order) were sufficient for them to correct the step on their own. This perceived usefulness is consistent with broader evidence that actionable, step-specific feedback can enhance students' ability to adjust their strategies in technology-enhanced environments (Barana et al., 2021; Koedinger et al., 2010). At the same time, students noted that the explanations did not always fully clarify deeper conceptual issues; when they were confused about why a particular transformation was appropriate or why an invariant should hold, they still turned to the lecturer for more detailed discussion. This balance between using XAI for procedural repair and relying on the lecturer for conceptual clarification is captured in the quotations under Theme 2 in Table 7.

Third, students framed XAI feedback as complementing, rather than replacing, lecturer feedback. Interviewees appreciated the immediacy and consistency of the XAI feedback. However, they emphasized that they still valued the lecturer's ability to explain concepts in multiple ways, connect tasks across the course, and provide motivational support. This echoes emerging evidence that explanations offered by automated scoring and analytics systems do not automatically increase student trust or motivation unless they are integrated into a broader instructional context (Banihashem et al., 2022). In our setting, students tended to trust the XAI scores when they aligned with their expectations. However, they relied on the lecturer to resolve discrepancies or elaborate on the conceptual meaning of the rubric indicators.

Overall, the qualitative findings align with XAI-ED frameworks that advocate human-centred design by tailoring explanations to stakeholder needs and embedding them into existing feedback practices (Khosravi et al., 2022). Students used XAI primarily as a fast, targeted diagnostic tool for their solutions. At the same time, lectures and in-class discussions remained the main arena for making sense of transformational geometry at a deeper level. Taken together, the three themes in Table 7 thus portray XAI feedback as a process-level support for locating and revising errors, embedded within a broader ecosystem of human-led conceptual explanation and motivational guidance.

Theoretical Implications: XAI Feedback As Process-Level Formative Feedback

Our quantitative and qualitative findings also have implications for how XAI-based step feedback can be interpreted within formative feedback theory. Hattie and Timperley's (2007) Model distinguishes four levels of feedback (task, process, self-regulation, and self) and highlights that feedback aimed at processes and self-regulation is often more powerful than feedback focused solely on task correctness. Our results suggest that, in this study, XAI-based step feedback operated primarily at the task and process levels: it pointed out whether specific steps were correct or incomplete (task), and in many cases indicated what needed to be changed in the procedure to align with rubric expectations (process), as reflected in overall learning gains in Table 5, Figures 2 and 7.

The process-level effects are most visible in the reductions in GT2 and GT5 error rates (Table 6, Figures 9–10) and in students' interview accounts of using GT indicators to locate problematic steps (Table 7). Several students explicitly mentioned GT2 and GT5 when describing how they rechecked the centre, line, or order of transformations in the XAI interface, mirroring the quantitative pattern of improvement on these indicators. This pattern resonates with recent meta-analytic work showing that feedback targeting processes and strategies tends to yield larger learning gains than purely outcome-focused information, especially in complex domains like mathematics (Wisniewski et al., 2020).

At the same time, the relatively modest, non-differential improvements in conceptual justification and verification (GT7–GT8) underscore the limits of the current XAI design for supporting self-regulation or higher-level conceptual change. This pattern is consistent with broader observations that current transformer-based language models remain relatively unrobust in evaluating rich, context-dependent mathematical arguments, especially when training data for such high-level reasoning steps are relatively sparse. The system signalled where problems occurred but did not provide the kind of explanation or prompting that would help students plan, monitor, and evaluate their own problem-solving at a metacognitive level. From an XAI-ED perspective, the study illustrates the benefits and trade-offs of aligning explanations with domain-specific rubrics. Khosravi et al.'s XAI-ED framework emphasises that explanations should be understandable for stakeholders, grounded in pedagogically meaningful constructs, and evaluated alongside learning outcomes and fairness (Khosravi et al., 2022). In our design, mapping explanations to GT1–GT8 indicators appears to have made model outputs more interpretable for both students and instructors and to have supported targeted process-level revisions, as evidenced by the fidelity metrics in Table 3.

However, as highlighted in recent XAI-in-education reviews, transparency at the task or feature level does not automatically translate into deeper understanding or trust; these outcomes depend on how explanations are orchestrated within teaching practices and how they interact with learners' prior knowledge (Liu, Pinto, & Paquette, 2024). Theoretically, our findings support a view of XAI-based assessment as a complementary actor in a human–AI feedback system rather than as a standalone feedback provider. Learning analytics research has shown that when teachers have access to fine-grained process data, they can better target instructional time to the concepts and tasks that most challenge students (Banihashem et al., 2022; Schwendimann et al., 2017). Similarly, XAI-based step scoring and explanations can take over the routine identification and signalling of procedural issues, freeing instructors to concentrate on designing activities and discussions that foster conceptual justification, proof, and self-regulation in transformational geometry. Rather than "bridging" cognitive constructs in a strong sense, our results indicate that XAI-based explanations, when aligned with a rubric like GT1–GT8, can help connect model decisions to observable aspects of students' solution processes in pedagogically actionable ways, but still require teacher mediation to achieve deeper conceptual and metacognitive goals.

Limitations and Future Work

Although the results are promising for the Transformational Geometry course studied here, several limitations should be noted. Firstly, the pedagogical phase involved only two intact classes in a single Mathematics Education programme at Universitas

Islam Malang. Although baseline checks and ANCOVA adjustment indicated broadly similar pre-test performance, residual class differences (e.g., preparation or classroom dynamics) cannot be ruled out. Future studies should replicate the design across more classes and sites to enable multilevel models and more robust subgroup analyses.

Secondly, the intervention was limited to one mathematical topic—transformational geometry in one curriculum. The course was delivered in an Islamic private university in East Java with a strong emphasis on teaching practice, so the findings may not fully transfer to institutions with different student populations, technological infrastructures, or assessment cultures. The XAI model was trained on step traces from this field and fitted to a GT1–GT8 aligned rubric. There is therefore little evidence as to whether the findings are generalisable to other areas of mathematics (e.g., algebra, calculus, statistics), to other levels of academic learning, or to institutions with varying cultural, technological, and assessment infrastructures. It may be significant to extend the method to other topics or contexts and to investigate further whether the same pattern of system-wide process-level improvement or learning gains occurs, and to what extent similar patterns of process-level improvement and learning gains emerge.

Thirdly, the current implementation of the XAI system operates entirely on textual and symbolic step traces and does not evaluate hand-drawn diagrams or dynamic constructions. This design choice was partly driven by the constraints of the LMS, where students entered their solutions in text boxes, and any sketches were produced on paper outside the system. As a result, when discrepancies arise between a student's written steps and their informal diagrams, only the written reasoning is reflected in the automated scores. In the present course, the lecturer could still inspect students' diagrams during class discussions, but from a measurement perspective, it remains an important limitation. Extending the approach to multimodal input for example, by combining step-level text models with recent vision–language architectures for handwritten mathematics—represents a promising direction for future work. Finally, the current generation of XAI feedback primarily focuses on providing task- and process-level information about students' solution steps. Even though this fits comfortably within the theoretical narrative that espouses the efficacy of process-level feedback, the interviews and rubric results indicate that higher-order conceptual justification and metacognitive regulation are, to a large extent, still human. Future work might investigate hybrid designs in which XAI-based step feedback is purposefully designed to integrate teacher-led discussion, peer-based collaboration, and guided prompts that address self-regulation and explore potential alternative explanations, thereby more effectively scaffolding conceptual thinking rather than merely signalling what to address.

▪ CONCLUSION

This study examined whether an explainable AI (XAI)–based system can validly and reliably score students' solution steps in transformational geometry, how faithful and fair its explanations are, and whether step-level XAI feedback can improve learning in an authentic university course. Across the validation analyses, the system approximated expert step scoring with agreement levels typically considered acceptable for educational assessment, while local explanation checks indicated that the highlighted features were meaningfully related to model predictions and showed no large performance disparities across gender or study programme. In the classroom quasi-experiment, students who

received XAI-based feedback achieved higher adjusted post-test scores than those who received conventional rubric-based feedback, with moderate learning gains over a short intervention. Process-level analyses indicated that these gains were concentrated on indicators related to parameter specification and composition of transformations, and interview data showed that students used the XAI interface to locate and correct specific steps rather than merely checking final answers. In our Transformational Geometry classes at Universitas Islam Malang, the combination of psychometric results, time-efficiency data, and classroom outcomes indicates that rubric-aligned XAI step assessment can provide scalable task- and process-level feedback without replacing the lecturer's role. These conclusions should be interpreted in light of the current XAI system, which evaluates only textual and symbolic steps, while the lecturer remains responsible for monitoring students' use of diagrams and for orchestrating discussions that address deeper conceptual understanding in transformational geometry.

At the same time, the work has important implications, limitations, and avenues for future research. In practice, the results point towards a human-in-the-loop configuration in which XAI handles routine, high-volume step scoring and provides immediate, rubric-aligned feedback on the technical aspects of students' solutions. At the same time, lecturers focus on facilitating conceptual discussion, addressing justification and verification indicators, and supporting students' motivation and self-regulation. The study is constrained by its small sample, two intact classes from a single institution, and focus on a single topic and short-term outcomes; as discussed in the Limitations and future work section, these factors call for cautious generalisation and motivate replication across additional cohorts, institutions, and mathematical domains. Future research should also explore how XAI explanations can be redesigned to better support conceptual justification and self-regulation, how teacher and peer perspectives can inform the orchestration of XAI feedback in classroom practice, and how more advanced psychometric and multilevel methods can be used to evaluate XAI-based assessment at scale. Within these boundaries, the present study contributes an integrated evaluation of XAI-based step assessment in mathematics. It illustrates how such systems can be embedded in formative assessment practices in ways that complement rather than replace human expertise.

▪ ACKNOWLEDGMENTS

The author would like to express sincere appreciation to the Institute for Research and Community Service (LPPM), Universitas Islam Malang, for their invaluable support and encouragement throughout the conduct of this research. Their guidance and assistance greatly contributed to the successful completion of this study.

▪ REFERENCES

Abazi Chaushi, B., Selimi, B., Chaushi, A., & Apostolova, M. (2023). Explainable artificial intelligence in education: a comprehensive review. In *Communications in Computer and Information Science*: *Vol. 1902. Explainable Artificial Intelligence (xAI, 2023)* (pp. 48–71). Springer. https://doi.org/10.1007/978-3-031-44067-0_3

Ada, T., & Kurtulus, A. (2010). Students' misconceptions and errors in transformation geometry. *International Journal of Mathematical Education in Science and Technology*, *41*(7), 901–909. https://doi.org/10.1080/0020739X.2010.486451

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). *Sanity checks for saliency maps*. Retrieved from https://papers.nips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Ballance, O. J. (2024). Sampling and randomisation in experimental and quasi-experimental CALL studies: Issues and recommendations for design, reporting, review, and interpretation. *ReCALL*, *36*(1), 58–71. https://doi.org/DOI: 10.1017/S0958344023000162

Banihashem, S. K., Mahroeian, H., Khosravi, H., Sadiq, S., & Gasevic, D. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, *37*, 100489. https://doi.org/10.1016/j.edurev.2022.100489

Barana, A., Marchisio, M., & Sacchet, M. (2021). Interactive feedback for learning mathematics in a digital learning environment. *Education Sciences*, *11*(6), 279. https://doi.org/10.3390/educsci11060279

Barredo Arrieta, A. (2024). Explainable AI (XAI) 2.0: A manifesto of open challenges and research directions. *Information Fusion*, *103*, 102224.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Clarke, V., & Braun, V. (2021). *Thematic analysis: A practical guide*. London, UK: SAGE.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

Elagha, N., & Pellegrino, J. W. (2024). Understanding error patterns in students' solutions to linear function problems to design learning interventions. *Learning and Instruction*, *92*, 101895. https://doi.org/10.1016/j.learninstruc.2024.101895

Fuchs, A., Gliwiński, M., Grageda, N., Spiering, R., Abbas, A. K., Appel, S., … Trzonkowski, P. (2018). Minimum information about t regulatory cells: a step toward reproducibility and standardization. *Frontiers in Immunology*, *8*, 1–14.

Gerke, O. (2020). Reporting standards for a Bland–Altman agreement analysis: A review of methodological reviews. *Diagnostics*, *10*(5), 334. https://doi.org/10.3390/diagnostics10050334

Gillard, D., Wright, D., McNally, A., Flaxman, P. E., McIntosh, R., & Honey, K. (2021). Acceptance & commitment therapy for school leaders' well-being: an initial feasibility study. *Educational Psychology in Practice*, *37*(1), 34–51. https://doi.org/10.1080/02667363.2020.1855120

Götz, S., & Gasteiger, H. (2022). Reflecting geometrical shapes: Approaches of primary students to reflection tasks. *Educational Studies in Mathematics*, *110*(2), 241–265.

Green, J. (2023). Primary students' experiences of formative feedback in mathematics. *Education Inquiry*, *14*(3), 285–305. https://doi.org/10.1080/20004508.2021.1995140

Hadjerrouit, S., & Nnagbo, C. I. (2022). Exploring Numbas formative feedback for teaching and learning mathematics: An affordance theory perspective. *Proceedings of the 18th International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA 2021)*. IADIS Press.

Hao, S., Pan, H., & Zhang, D. (2025). A process-oriented approach to assessing high school students' mathematical problem-solving competence: insights from multidimensional eye-tracking analysis. *Education Sciences*, *15*(6), 761. https://doi.org/10.3390/educsci15060761

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hedges, L. V, Tipton, E., Zejnullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *British Journal of Mathematical and Statistical Psychology*, *76*(2), 259–282. https://doi.org/10.1111/bmsp.12296

Herbert, S., Vale, C., White, P., & Bragg, L. A. (2022). Engagement with a formative assessment rubric: A case of mathematical reasoning. *International Journal of Educational Research*, *111*, 101899. https://doi.org/10.1016/j.ijer.2021.101899

Hirose, M., & Creswell, John W. (2022). Applying Core Quality Criteria of Mixed Methods Research to an Empirical Study. *Journal of Mixed Methods Research*, *17*(1), 12–28. https://doi.org/10.1177/15586898221086346

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., & Buckingham Shum, S. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*(4), 629–653. https://doi.org/10.1007/s40593-021-00239-1

Hontvedt, M., Prøitz, T. S., & Silseth, K. (2023). Collaborative display of competence: A case study of process-oriented video-based assessment in schools. *Teaching and Teacher Education*, *121*, 103948. https://doi.org/10.1016/j.tate.2022.103948

Hoth, J., Larrain, M., & Kaiser, G. (2022). Identifying and dealing with student errors in the mathematics classroom: Cognitive and motivational requirements. *Frontiers in Psychology*, *13*, 1057730. https://doi.org/10.3389/fpsyg.2022.1057730

Khalil, M., Prinsloo, P., & Slade, S. (2023). Fairness, trust, transparency, equity, and responsibility in learning analytics (Editorial). *Journal of Learning Analytics*, *10*(1), 1–7.

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Gašević, D., Kay, J., … Tsai, Y.-S. (2022). Explainable artificial intelligence in education. *Computers & Education: Artificial Intelligence*, *3*. https://doi.org/10.1016/j.caeai.2022.100074

Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining* (pp. 43–56). CRC Press. https://doi.org/10.1201/b10274-10

Koo, T. K., & Li, M. Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

https://doi.org/10.1177/1948550617697177

Li, M., Gao, Q., & Yu, T. (2023). Kappa statistic considerations in evaluating inter-rater reliability between two raters: Which, when, and context matters. *BMC Cancer*, *23*, 799. https://doi.org/10.1186/s12885-023-11325-z

Liu, Q., Pinto, J. D., & Paquette, L. (2024). *Applications of explainable AI (XAI) in Education BT - trust and inclusion in ai-mediated education: where human learning meets learning machines* (D. Kourkoulou, A.-O. (Olnancy) Tzirides, B. Cope, & M. Kalantzis, Eds.). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64487-0_5

Lopes, M. N. (2024). An overview of the empirical evaluation of Explainable AI (XAI): A systematic review. *Applied Sciences*, *14*(23), 11288.

Maskos, K., Schulz, A., Oeksuez, S. S., & Rakoczy, K. (2025). Formative assessment in mathematics education: A systematic review. *ZDM–Mathematics Education*, *57*(4), 679–693. https://doi.org/10.1007/s11858-025-01696-x

Mathaba, P. N., Bayaga, A., Tîrnovan, D., & Bossé, M. J. (2024). Error analysis in algebra learning: Exploring misconceptions and cognitive levels. *Journal on Mathematics Education*, *15*(2), 575–592. https://doi.org/10.22342/jme.v15i2.pp575-592

Mbusi, N., & Luneta, K. (2023). Implementation of an intervention program to enhance student teachers' active learning in transformation geometry. *SAGE Open*, *13*(2).

Mbusi, N. P., & Luneta, K. (2021). Mapping pre-service teachers' faulty reasoning in geometric translations to the design of Van Hiele phase-based instruction. *South African Journal of Childhood Education*, *11*(1), a871. https://doi.org/10.4102/sajce.v11i1.871

Miró-Nicolau, M., Jaume-i-Capó, A., & Moyà-Alcover, G. (2024). A comprehensive study on fidelity metrics for XAI. *Information Processing & Management*, *61*, 103988.

Ndlovu, M. (2022). Teachers' challenges in implementing philosophical approaches to mathematics education. *Pythagoras*, *43*(1), 1–10. https://doi.org/10.4102/pythagoras.v43i1.647

Ndungo, I. (2024). *A qualitative investigation on learners' experiences and understanding of transformation geometry with van hiele phased instruction and technology-enhanced van hiele phased instruction*.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. *Proceedings of BMVC*, Vol. 151.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of NAACL: Demonstrations*, pp. 97–101. https://doi.org/10.18653/v1/N16-3020

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Shirvani Boroujeni, M., Holzer, A., … Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. https://doi.org/10.1109/TLT.2016.2599522

Shimizu, Y., & Kang, H. (2025). Research on classroom practice and students' errors in mathematics education: A scoping review of recent developments for 2018–2023. *ZDM – Mathematics Education, 57*, 695–710. https://doi.org/10.1007/s11858-025-01704-0

Söderström, S., & Palm, T. (2024). Feedback in mathematics education research: A systematic literature review. *Research in Mathematics Education*.

St. Goar, J., & Lai, Y. (2022). Designing Activities to Support Prospective High School Teachers' Proofs of Congruence from a Transformation Perspective. *PRIMUS*, *32*(7), 827–842. https://doi.org/10.1080/10511970.2021.1940403

Sundararajan A.; Yan, Q., M. T. (2017). Axiomatic attribution for deep networks. *Proceedings of ICML*, *70*, 3319–3328. Retrieved from https://proceedings.mlr.press/v70/sundararajan17a.html

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328. https://doi.org/10.48550/arXiv.1703.01365

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2024). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*, *29*(5), 967–979. https://doi.org/10.1037/met0000516

Tornqvist, M., Mahamud, M., Méndez Guzmán, E., & Farazouli, A. (2023). ExASAG: Explainable framework for automatic short answer grading. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA, 2023)*, 361–371. https://doi.org/10.18653/v1/2023.bea-1.29

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zorn, K., Larkin, K., & Grootenboer, P. (2022). Student perspectives of engagement in mathematics. In N. Fitzallen, C. Murphy, V. Hatisaru, & N. Maher (Eds.), *Mathematical confluences and journeys: Proceedings of the 44th Annual Conference of the Mathematics Education Research Group of Australasia (MERGA)* (pp. 570–577). Launceston, Australia: MERGA.

Zumba-Zúñiga, M.F., Ríos-Zaruma, J., Pardo-Cueva, M., & Chamba-Rueda, L. (2021). Impact of information and communication technologies in Higher Education Institutions in times of COVID-19 : A look from collaborative work and study modality. *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI52073.2021.9476642