Modeling the Role of Graphical and Textual Complexity in Geometry Problem-Solving Using SEM

Farshid Farzan[1],Canwen Wang[2], Rachel L Ankney[3] and Paulo Carvalho[4]

[1]Cognitive Group, Psychology Department, University of Memphis, USA

[2]Human Computer Interaction Institute, Carnegie Mellon University, USA

[3]Psychology Department, University of Memphis, USA

[4]Human Computer Interaction Institute, Carnegie Mellon University, USA

ffarzan@memphis.edu
canwenw@andrew.cmu.edu
rlankney@memphis.edu
pcarvalh@cs.cmu.edu

Abstract: Learning geometry is vital for K–12 students, fostering spatial reasoning and problem-solving skills crucial in STEM fields. However, its abstract nature, reliance on visualization, deductive reasoning, and specific vocabulary make it challenging. Cognitive Load Theory explains how excessive graphical and textual complexity can overwhelm working memory, impeding learning. The revised Cognitive Load Theory framework highlights the importance of directing cognitive effort toward meaningful schema construction (germane processing) rather than being hindered by extraneous load. This study uses Structural Equation Modeling (SEM) to examine how structural complexity (graphical and textual complexity) affects student performance in geometry learning, modeled as a latent variable measured by accuracy, latency, and first correct attempt. Forty geometry questions exam answered by 59 students from the 1996–97 Cognitive Tutor dataset (DataShop) were analyzed to examine how graphical and textual complexity influenced student performance. Each question systematically varied in complexity, and student performance was measured using accuracy rate, latency, and first correct attempt, well-established metrics in educational research. Student Performance is conceptualized as a latent variable measured by Accuracy Rate, Latency, and First Correct Attempt. Graphical Complexity and Textual Complexity serve as observed variables. Employing a two-phase SEM approach, the study evaluates both measurement and structural models. The findings reveal that Graphical Complexity and Textual Complexity significantly and negatively influence student performance overall, with unstandardized regression coefficients of -.132 and -.142, respectively, and explain 40.9% of the variance in SP. However, contrary to the initial understanding that it seems the higher complexity may lead to initial wrong answers, higher levels of Graphical Complexity and Textual Complexity were associated with increased First Correct Attempt, suggesting that complexity may promote deeper cognitive engagement in some contexts aligned with Cognitive Load Theory, highlighting both the detrimental effects of complexity on efficiency and the potential benefits for accuracy in problem-solving.

Keywords: Geometry Problem-Solving, Structural Complexity, Student Performance, Cognitive Load Theory, Structural Equation Modeling

1. Introduction

Learning geometry is useful for K-12 students, as it develops critical spatial reasoning, problem-solving, and foundational mathematical skills that apply across disciplines and real-world contexts (Mix & Battista, 2018) and enhance their spatial visualization and logical thinking, key components in STEM fields (Sinclair & Bruce, 2014). Learning geometry presents unique challenges for K-12 students, largely due to its abstract nature and the need for strong spatial reasoning skills (Dhlamini et al, 2019). Many students struggle with visualizing and mentally manipulating shapes in two- and three-dimensional spaces, which is essential for understanding concepts like symmetry, transformations, and spatial relationships (Clements & Battista, 1992). Additionally, geometry often requires a level of deductive reasoning and logical proof construction that can be challenging for younger learners, as these skills are not as heavily emphasized in other areas of math (Hiele, 1986). Language also plays a role (Darke, 1982); geometry introduces specific terms and definitions that may confuse students, especially if they lack prior exposure to geometric vocabulary. Moreover, the interplay of graphical and textual complexity (TC) in problem-solving environments adds further cognitive load (Lin & Lin, 2014), potentially impairing student performance (SP). Cognitive Load Theory (CLT) (Sweller, 1988) provides a theoretical framework for this study, positing that excessive visual or textual demands can overwhelm working memory and hinder learning outcomes.

CLT (Sweller, 1988) emphasizes that excessive visual or textual demands can overwhelm working memory and hinder learning. Initially, CLT categorized cognitive load into intrinsic, extraneous, and germane load, with the latter referring to cognitive effort for schema formation (Sweller, 1988). However, the 2019 revision (Sweller et

al, 2019) removed germane load as a distinct category, redefining it as germane processing, a qualitative aspect of learning rather than a separate cognitive load component. Empirical studies (Husni et al, 2022) suggest that learning efficiency depends on learners' engagement in meaningful schema construction. In geometry problem-solving, complexity arises from various components (Bobis et al, 1993), which can either impose extraneous load or contribute to germane processing depending on how students interact with the problem. High graphical complexity (GC) may increase cognitive burden but can also enhance engagement by encouraging integration of visual and conceptual understanding (Bobis et al, 1993; Lin & Lin, 2014). Similarly, textual complexity, which involves the structure, length, and syntactic complexity of problem statements, can be assumed informational elements (Lin & Lin, 2014), may require additional working memory resources, but can also encourage students to engage in more deliberate cognitive processing (Seufert, 2018). According to the revised CLT framework, learning occurs most effectively when cognitive resources are directed toward germane processing rather than being overwhelmed by excessive extraneous load (Sweller et al, 2019). This perspective underscores the importance of understanding how problem complexity influences cognitive engagement, particularly in geometry (Clements & Battista, 1992; Dhlamini et al, 2019), where effective problem-solving relies on the interplay between visual representation, conceptual reasoning, and textual interpretation.

This study models SP as a latent construct measured by Accuracy Rate (AR), Latency (LT), and First Correct Attempt (FCA), with GC and TC as key predictors. Following CLT, structural Equation modeling (SEM) analysis evaluates how problem complexity is associated with SP.

2. Method

2.1 Participants and Test Items

The study utilized 40 geometry questions derived from the Geometry Area (1996-97) dataset, accessed via DataShop (Koedinger et al, 2010). This database is widely utilized in educational research and is recognized for its robust design within the Cognitive Tutor framework, ensuring the quality and relevance of the questions for analyzing SP in geometry. The data for this study were collected from the area unit of a Geometry course conducted during the 1996–1997 school year. Specifically, the dataset was derived from activities completed on February 1, 1996, using the Cognitive Tutor 1996 system, which was designed to support student learning in mathematics. The subset used in this study includes interaction data from 59 students, encompassing a total of 3591 problem-solving steps recorded during their engagement with the geometry problems. This dataset focuses on the area subject within Geometry and captures SP metrics from that period. Unfortunately, additional contextual information, such as the names of participating schools or specific student demographics, is not available in the dataset. The absence of such data limits the ability to generalize findings to specific educational settings but does not impact the core analysis of SP and its relationship to problem complexity. These questions, designed within the framework of the Cognitive Tutor system, are commonly used in educational research to measure SP in geometry (Stamper & Koedinger, 2011). Each question varies systematically in GC and TC, allowing for an analysis of their associations with three key SP metrics: AR (proportion of incorrect responses), LT (time to complete each question), and FCA (success on the first attempt). These metrics are well-established in educational research as indicators of SP, particularly in studies examining cognitive and problem-solving processes in mathematics (Clements & Battista, 1992; Jones & Tzekaki, 2016).

In this study, TC refers to the complexity of the textual content in geometry problems, evaluated using a range between 1 to 3 based on question text length (Table 1).

Table 1: Rubric for geometry problem TC

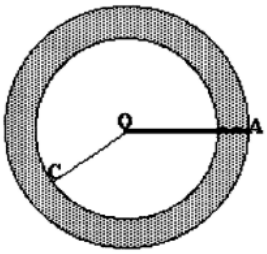
Level (Value)	Criteria
Low (1)	Question with fewer than 30 words
Medium (2)	Question with 30 to 60 words
High (3)	Question with more than 60 words

GC represents the complexity of visual elements, also evaluated using a range between 1 to 3 based on the number and arrangement of shapes, required spatial reasoning, and graphical detail (Table 2).

Table 2: Rubric for geometry problem GC

Level (Value)	Criteria
Low	One plane shape
Medium	Plane shape with one additional component
High	Plane shape with more than one additional component

These categorizations were designed to align with CLT (Sweller, 1988), which suggests that increases in TC or GC impose greater cognitive demands on learners. A sample of these questions has been shown in Figure 1. The rubric criteria for GC and TC were developed through a combination of theory-driven reasoning and practical analysis of item features, consistent with CLT's focus on surface- and structure-level complexity (Sweller, 1988; Sweller et al, 2019). For TC, word count was used as a proxy for syntactic and semantic load, which aligns with approaches in prior educational psychology and text complexity research (Lin & Lin, 2014). GC levels were informed by visual inspection of problem diagrams, focusing on the number and configuration of visual components and the spatial reasoning required elements previously shown to affect cognitive demand in geometry problem-solving tasks (Bobis et al, 1993; Dhlamini et al, 2019). While simplified, these rubrics offer a structured and replicable way to classify complexity across a standardized item set and serve as a basis for exploratory modeling.



Problem Statement

In the figure, the shaded area is 60.75π square inches. If OC = 4.5 inches, find OA.

$\pi = 3.1416$.

Figure 1: Sample problem with TC = 1 and GC = 2

This study examines SP as a latent variable measured by three key indicators: AR, representing the proportion of correct responses (1 – Error Rate); LT, the time taken to complete problems; and FCA, the number of correct answers on the first attempt. GC and TC are the observed variables representing problem complexity, defined by visual elements (e.g., shape arrangement) and textual characteristics (e.g., sentence structure and length). The conceptual framework integrates measurement and structural models, hypothesizing that SP is associated with GC and TC, with a correlation between the two complexities to account for shared variance. In the structural equation model (SEM), GC and TC serve as exogenous predictors, while AR, LT, and FCA function as endogenous effect indicators of SP. These indicators capture distinct yet interrelated dimensions of SP: accuracy (AR), efficiency (LT), and mastery (FCA), ensuring the construct's validity. SP is standardized through Unit Loading Identification (ULI), fixing one path variance to unity for consistent interpretation of relationships. Residual variance accounts for unexplained variability, preventing measurement error and external influences from inflating relationships. For latent variables like SP, it captures unmeasured influences, while for observed variables like GC and TC, it accounts for noise, enhancing model validity and reflecting real-world complexity.

No control variables were included, as GC and TC are assumed to capture the primary cognitive demands affecting SP. Furthermore, the dataset does not include comprehensive information on participant demographics or prior knowledge, making it challenging to incorporate such variables reliably without introducing additional assumptions. While the absence of control variables may limit the ability to explore moderating influences, this choice ensures a clear and unbiased evaluation of the hypothesized relationships.

2.2 Two-Phase SEM Approach

To evaluate relationships in this study, the SEM model was analysed in two phases: the measurement phase, defining SP, and the structural phase, assessing the direct predictive effects of complexity factors (Brown, 2015).

In the measurement phase, SP is defined as a latent variable measured by its three indicators (AR, LT, and FCA). The observed variables GC and TC are included in the model with a bidirectional arrow between them, representing their correlation. In this phase, no causal or predictive relationships are assumed; instead, the focus is on confirming the validity and reliability of SP as a latent construct and assessing the covariation between GC and TC (Figure 1). In the structural phase, an improved model introduced to treat GC and TC as predictors, directly modeling their relationships with SP. This allows the model to test the direct effects of GC and TC on SP and assess how each complexity factor independently impacts SP.

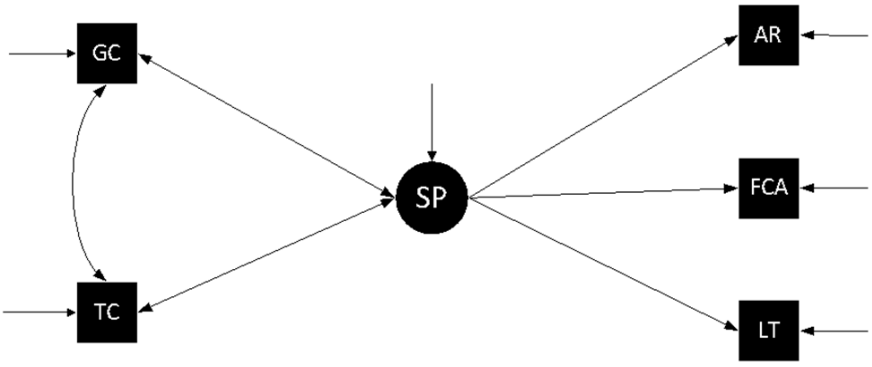


Figure 2: Measurement Phase Model Result

The dataset met all SEM assumptions with no violations in normality, linearity, or multicollinearity. Skewness, kurtosis, and VIF values were within acceptable limits, and no missing data or outliers were found, ensuring suitability for analysis. The measurement model, estimated using Maximum Likelihood (ML) in Mplus Version 8, converged successfully without issues, meeting the convergence criterion (.00005). The information matrix was positive definite, with no warnings regarding parameter identification, non-positive-definite covariance matrices, or linear dependencies among parameters. Furthermore, no offending estimates, such as negative error variances or Heywood cases, were identified.

3. Result

Descriptive statistics for the measured variables provide an overview of their distribution and characteristics. GC *(M = 2.20, SD =.71)*, and TC *(M = 1.93, SD = .85).* The indicators of the latent variable SP demonstrated the following properties: AR *(M = 2.05, SD = .65)*, LT *(M = .50, SD = .69)*, and FCA *(M = .73, SD = .76)*. (All indicators have been normalized to a range of 0 to 3.)

3.1 Model Fitting Analysis

In assessing model fit for this study, multiple indices were used to evaluate the adequacy of both the measurement and structural components. The chi-square (χ^2) test was included as an overall measure of global fit, with *p>.05* indicating acceptable model fit (Kline, 2023). Additional global fit indices included the Root Mean Square Error of Approximation (RMSEA), with values ≤ .05 indicating a good fit (Hu & Bentler, 1999), and the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI), where values ≥ .95 represent a good fit. The Standardized Root Mean Square Residual (SRMR), with values ≤ .08, was also used to evaluate fit (Hu & Bentler, 1999). To assess local fit, the analysis will focus on inspecting residuals to ensure the model accurately represents the data. Residuals will be evaluated by interpreting standardized residuals as z-scores, with values exceeding |1.96| (correlation residuals, .1) indicating potential misfit. For the two-step SEM process, model comparisons will be conducted using the chi-square difference test for nested models, allowing for a rigorous evaluation of model fit and theoretical refinement (Hancock & Mueller, 2019). Both the measurement and structural models were estimated using the Maximum Likelihood (ML) method in Mplus Version 8.

3.2 Measurement Model

The measurement model demonstrated an acceptable global fit to the data. The chi-square test yielded a value of $\chi^2(5) = 8.208$, $p = .145$, indicating no significant misfit. The RMSEA was .127, with a 90% confidence interval of .000 – .276, suggesting that the model is plausible, particularly given that the lower bound includes zero. The Comparative Fit Index (CFI) was .956, exceeding the threshold of .90, reflecting good comparative fit. Additionally, the Standardized Root Mean Square Residual (SRMR) value was .049, indicating minimal

discrepancies between observed and predicted values. The local fit assessment for the measurement model was satisfactory. Standardized residuals for the covariances were within acceptable limits, with none exceeding |1.96| and residuals for correlation with none exceeding |.1|, indicating that the covariances were adequately represented. The residual variances for the observed variables AR, LT, and FCA were statistically significant, confirming that the indicators contributed meaningfully to the measurement of the latent variable SP.

3.3 Structural Model

The structural model, which includes regression paths from GC and TC to SP, also exhibited good global fit. The chi-square test resulted in a value of $\chi^2(5) = 8.208$, $p = .145$, indicating no significant misfit. The RMSEA was .127, with a 90% confidence interval of .000 – .276, suggesting that the model is plausible, particularly given that the lower bound includes zero. The Comparative Fit Index (CFI) was .957, exceeding the threshold of .90, reflecting good comparative fit. Additionally, the Standardized Root Mean Square Residual (SRMR) value was .049, indicating minimal discrepancies between observed and predicted values. The local fit assessment for the measurement model was satisfactory. Standardized residuals for the covariances were within acceptable limits, with none exceeding |1.96| and residuals for correlation with none exceeding |.1|, indicating that the covariances were adequately represented. Residual variances for all observed variables were significant, supporting the proper specification of regression paths and the overall structural relationships. These results confirm that the structural model captures the hypothesized associations between GC and TC on SP effectively and is well-suited for testing the research hypotheses.

3.4 Model Comparisons

The measurement model yielded mixed fit indicators: $\chi^2(5) = 8.208$, $p = .145$, CFI = .956, and SRMR = .049, all of which suggest acceptable to good fit. However, the RMSEA of .127 (90% CI: .000–.276) exceeds commonly accepted thresholds, indicating potential local misfit. Given the non-significant chi-square and favourable CFI and SRMR values, the overall model fit is considered adequate, though interpretation should be made with caution due to the elevated RMSEA. The structural model, which added regression paths from GC and TC to SP, yielded identical fit indices: a chi-square value of 8.208 with 5 degrees of freedom, an RMSEA of .127 *(90% CI: .000–.276)*, a CFI of .957, and an SRMR of .049. Given the identical chi-square values and degrees of freedom, the measurement and structural models are statistically equivalent. This lack of change in fit reflects the fact that the two models cannot be statistically distinguished based on the current data. Despite this equivalence, the inclusion of the structural relationships is theoretically justified, as it allows for the examination of the direct relationships between GC and TC on SP. While the results should be interpreted cautiously, the structural model aligns with the theoretical framework being investigated. Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were also consistent across models, further supporting the appropriateness of the structural model without suggesting overfitting. Overall, the structural model was retained due to its theoretical relevance, despite no significant statistical difference with the measurement model in terms of fit indices.

3.5 Latent Factor Validity and Reliability

Validity and reliability analyses confirmed the appropriateness of SP as a latent construct. *Average Variance Extracted (AVE) = .610*, exceeding the .50 threshold, supports construct validity by explaining most indicator variance. *Composite Reliability (CR) = .650,* though below the ideal .70, indicates moderate reliability, suggesting room for measurement refinement. Overall, these results validate SP for further analysis.

3.6 Parameter Estimates and Variance

The structural model was evaluated using unstandardized parameter estimates to assess the relationships between latent and observed variables (Figure 2). The latent variable SP was measured by three observed indicators. For AR, the unstandardized parameter was fixed at 1.000 to scale the latent construct (ULI constraint). For LT, the unstandardized parameter was −2.453 *(SE = 1.034, p = .018)*, indicating that a one-unit increase in SP corresponds to a 2.453 decrease in LT. For FCA, the unstandardized parameter was −3.213 *(SE = 1.313, p = .014)*, suggesting that a one-unit increase in SP corresponds to a 3.213 decrease in FCA.

The structural regression paths indicated significant relationships between the predictors and SP. GC was a significant predictor with an unstandardized parameter of −.132 *(SE=.067, p = .050)*. This suggests that a one-unit increase in GC is associated with a .132 decrease in SP. Similarly, TC significantly predicted SP with an

unstandardized parameter of −.142 *(SE=.067, p = .035),* indicating that a one-unit increase in TC corresponds to a .142 decrease in SP. The covariance between GC and TC was not statistically significant *(p = .917).*

The variance explained (R2) for each endogenous variable demonstrated the contribution of the predictors to the latent construct. For SP, *R2=.409 (SE=.120, p = .001)*, indicating that GC and TC collectively explain 40.9% of the variance in SP. The variance explained for each indicator of SP reflects how well the latent construct predicts the observed variables. LT *(R2=0.702, SE=0.079, p<0.001)* demonstrates a strong relationship with SP, indicating that 70.2% of its variance is explained by the latent construct. Similarly, FCA *(R2=0.998, SE=0.000, p<0.001)* shows an almost perfect alignment with SP, as 99.8% of its variance is accounted for. In contrast, AR *(R2=0.130, SE=0.099, p=0.189)* contributes less significantly, with only 13.0% of its variance explained by SP. These findings suggest that LT and FCA are more strongly tied to the latent construct than AR.

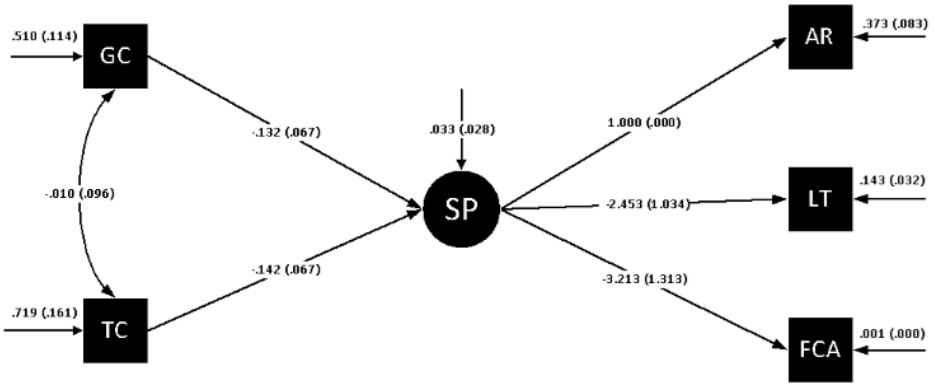


Figure 3: Full SEM for SP in geometry learning

4. Discussion

This study employed a two-phase SEM approach to investigate the relationships between GC, TC, and SP. The latent construct SP, measured through Accuracy Rate (AR), Latency (LT), and First Correct Attempt (FCA), demonstrated adequate validity and moderate reliability. The structural model revealed statistically significant negative associations between both GC and TC with SP, aligning with the hypothesis that increased complexity may correspond to decreased SP.

While the measurement and structural models showed statistically equivalent fit, the structural model remains theoretically informative. Its inclusion supports the exploration of how complexity metrics relate to SP outcomes in geometry problem-solving. However, this equivalence also highlights the importance of future methodological refinements, including the integration of additional variables, more recent datasets, or longitudinal and experimental designs that can support stronger causal inferences.

These findings are interpreted through the lens of Cognitive Load Theory (CLT). The negative associations observed between both TC and GC with SP are consistent with the notion that increased complexity imposes extraneous cognitive load, potentially overwhelming working memory and interfering with task SP. Specifically, TC likely contributes to extraneous load through longer problem statements, syntactic density, or unnecessary wording that may divert cognitive resources away from schema acquisition or conceptual reasoning. Higher visual complexity may also impose extraneous load, particularly when diagrams are cluttered or irrelevant to the learning objective.

However, according to the results of this study, it seems that when the geometry evaluation materials are designed effectively, structural complexity may contribute to germane processing, supporting schema construction by engaging students in spatial reasoning and conceptual integration (Sweller et al, 2019). This theoretical interpretation aligns with the observed positive relationship between complexity and First Correct Attempt (FCA), suggesting that higher complexity may prompt learners to allocate greater cognitive effort and engage more deeply with problem-solving processes. Another possible explanation for the positive link between complexity and FCA is that students may have responded to more complex problems with increased attentional effort, especially when the problem format encouraged active engagement.

However, it is important to consider that certain problem types or student characteristics (e.g., prior knowledge, motivation) may moderate this relationship. Future research should test this possibility by examining whether the effect of complexity on FCA varies by problem type or levels of learner involvement, using larger and more

diverse datasets. Additionally, some high-complexity problems may have incidentally included features, such as clearer spatial cues or better-organized information, that offset their cognitive demands. It is also possible that individual differences in prior knowledge, spatial ability, or mathematical experience moderated how students interacted with complex problems. Learners with stronger background skills may have been better able to manage cognitive demands and capitalize on opportunities for successful problem-solving on the first attempt.

5. Conclusion

By modeling SP as a latent construct composed of accuracy, efficiency, and mastery, this study offers a nuanced understanding of how different forms of problem complexity relate to geometry problem-solving behaviours. The use of a two-phase SEM approach allows for the exploration of these relationships in a theoretically grounded yet data-driven manner.

These findings have important implications for the application of Cognitive Load Theory (CLT) in educational design. The results suggest that increased graphical and textual complexity is not uniformly associated with decreased SP. While higher complexity levels may impose additional cognitive demands, they can also be linked to deeper cognitive engagement and more deliberate problem-solving strategies, particularly in cases where students succeed on their first attempt. This highlights the dual nature of complexity: it can either overload working memory (extraneous load) or support meaningful schema construction (germane processing), depending on the design of instructional materials and learners' cognitive readiness. Rather than minimizing difficulty outright, educators and curriculum designers should aim to optimize complexity levels to promote germane processing while minimizing unnecessary cognitive burden. In doing so, they can create learning environments that balance challenge and clarity, enhancing both engagement and SP.

Future research should expand on these findings by incorporating learner-level characteristics such as prior knowledge, cognitive ability, and motivational factors. Identifying which students benefit most from high-complexity tasks could inform the development of adaptive learning systems that personalize complexity based on learner profiles. Additionally, extending this research with more recent datasets and experimental designs will further strengthen the generalizability and practical applicability of these findings in contemporary educational settings.

6. Limitations

While the dataset originates from 1996–1997, its use is still valuable because it was collected within a well-established intelligent tutoring system (Cognitive Tutor) whose structure and instructional design principles continue to inform modern adaptive learning systems. Moreover, the underlying cognitive principles of geometry problem-solving and cognitive load remain highly relevant across decades. However, we recognize that technological tools and curricula have evolved, and we strongly encourage future research to replicate and extend these findings using more recent datasets to validate the robustness of the observed relationships.

It is important to recognize that these results are based on cross-sectional data, which limits the ability to establish definitive causal relationships. While Structural Equation Modeling (SEM) is a powerful tool for testing theoretically informed models, acceptable fit indices only suggest that the model is one plausible explanation for the observed data. Alternative explanations, such as unmeasured confounding variables or bidirectional influences, could also account for the observed associations. Future research should address these limitations by employing longitudinal or experimental designs that allow for robust testing of temporal and causal mechanisms. Also, future work may refine these categories using more granular linguistic and visual analysis tools or machine-coded complexity metrics. Such approaches could enhance objectivity and replicability in complexity coding, reducing potential bias introduced by manual rubric-based classification.

Additionally, the dataset does not include detailed participant information such as demographics or prior knowledge measures. While this focused sample enables close examination of problem-solving behaviour in a real-world educational setting, the absence of prior knowledge indicators limits our ability to assess how individual differences in mathematical background may have influenced SP or moderated the effects of complexity. Future research should consider incorporating background measures or pre-assessments to better capture learner variability and cognitive preparedness.

Ethics Declaration

This study involved secondary analysis of de-identified log data from student interactions with the Cognitive Tutor system, originally collected in an educational setting during the 1996–1997 school year. No personal, demographic, or sensitive information was included in the dataset, and no direct contact with participants occurred. As the dataset is fully anonymized and publicly accessible via DataShop, formal ethical clearance was not required. All procedures complied with institutional and field-specific guidelines for ethical use of archival educational data involving minors.

AI Declaration

During the preparation of this manuscript, the authors used GPT-4 for grammar and clarity checks. No content was generated by the AI tool beyond surface-level editorial suggestions. All substantive content, analysis, and interpretations were developed by the authors, who take full responsibility for the final work.

References

Bobis, J., Sweller, J., and Cooper, M. (1993). Cognitive load effects in a primary-school geometry task. Learning and Instruction, 3(1), pp 1–21. https://doi.org/10.1016/S0959-4752(09)80002-9

Brown, T. A. (2015). Confirmatory factor analysis for applied research. Guilford Publications.

Clements, D. H., and Battista, M. T. (1992). Geometry and spatial reasoning. Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics, pp 420-464.

Darke, I. (1982). A review of research related to the topological primacy thesis. Educational Studies in Mathematics, 13, pp 119–142. https://doi.org/10.1007/BF00460707

Dhlamini, Z. B., Chuene, K., Masha, K., and Kibirige, I. (2019). Exploring grade nine geometry spatial mathematical reasoning in the South African Annual National Assessment. Eurasia Journal of Mathematics, Science and Technology Education, 15(11). https://doi.org/10.29333/ejmste/105481

Hancock, G. R., and Mueller, R. O. (Eds.). (2019). The reviewer's guide to quantitative methods in the social sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203861554

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), pp 1–55. https://doi.org/10.1080/10705519909540118

Husni, N. A., Jumaat, N., and Tasir, Z. (2022). Investigating student's cognitive engagement, motivation, and cognitive retention in learning management system. International Journal of Emerging Technologies in Learning, 17(9), pp 184–200. https://doi.org/10.3991/ijet.v17i09.29727

Jones, K., and Tzekaki, M. (2016). Research on the teaching and learning of geometry. The second handbook of research on the psychology of mathematics education: The journey continues, pp 109-149. https://doi.org/10.1007/978-94-6300-561-6_4

Kline, R. B. (2023). Principles and practice of structural equation modeling (5th ed.). Guilford Publications.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining, 43, pp 43-56.

Lin, J. J. H., and Lin, S. S. (2014). Cognitive load for configuration comprehension in computer-supported geometry problem solving: An eye movement perspective. International Journal of Science and Mathematics Education, 12, pp 605–627. https://doi.org/10.1007/s10763-013-9479-8

Mix, K. S., and Battista, M. T. (2018). Visualizing Mathematics: The Role of Spatial Reasoning in Mathematical Thought. Research in Mathematics Education. Research in Mathematics Education. https://doi.org/10.1007/978-3-319-98767-5

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. Educational Research Review, 24, pp 116–129. https://doi.org/10.1016/j.edurev.2018.03.004

Sinclair, N., Bruce, C., Caswell, B., Lissa, D., Amour, B., Davis, M., Drefs, I., Elia, T., Flynn, K., Francis, D., Hallowell, Z., Hawes, H., Kaur, A., Mamolo, L., McGarvey, J., Moss, S., Naqvi, O.-L., Ng, Y., Okamoto, A., and Hawes, Z. (2014). Spatial reasoning for young learners. https://doi.org/10.13140/2.1.2543.5521

Stamper, J. C., and Koedinger, K. R. (2011). Human-machine student model discovery and improvement using DataShop. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), Proceedings of the 15th International Conference on Artificial Intelligence in Education (AIED 2011), pp 353–360. Springer. https://doi.org/10.1007/978-3-642-21869-9_46

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), pp 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31, pp 261–292.

Van Hiele, P. M. (1986). Structure and insight: A theory of mathematics education. Academic Press.