ACTA SCIENTIAE

ISSN: 2178-7727
DOI: 10.64856/acta.scientiae.8363

Process Over Product: Formal Logic, Sudoku Puzzles, and the Development of Reasoning in Student Teachers

Raúl Martínez-Bohórquez[a]
Lina Melo[b]
Lucía Bautista-Bárcena[c]

[a] Universidad de Extremadura, Escuela de Ingenierías Industriales, Departamento de Matemáticas, Badajoz, España.
[b] Universidad de Extremadura, Facultad de Educación y Psicología, Departamento de Didáctica de las Ciencias Experimentales y las Matemáticas, Badajoz, España.
[c] Universidad de Extremadura, Facultad de Ciencias, Departamento de Matemáticas, Badajoz, España.

ABSTRACT

Background: A critical disconnect often exists between the formal logical reasoning taught to student teachers and the applied, practical problem-solving skills required in a classroom. This study explores this "far transfer" problem using complex logic puzzles. **Objectives:** This research aimed to investigate the relationship between student teachers' formal logical reasoning skills, as measured by the Test of Logical Thinking (TOLT), and their performance, efficiency, and metacognitive strategies when solving variant Sudoku puzzles. **Design:** The study employed a descriptive correlational design to analyze the relationship between participants' cognitive skills and their puzzle-solving behaviors. **Setting and Participants:** The study involved 69 volunteer pre-service teachers from the Degree in Primary Education program at the Universidad de Extremadura, Spain, during the 2024/2025 academic year. **Data collection and analysis:** Data were gathered using the TOLT, puzzle-solving metrics (resolution rates, time, retries), and reflective questionnaires. The analysis involved Spearman's rank correlation and thematic analysis of qualitative responses. **Results:** TOLT scores did not significantly predict success in solving the puzzles ($\rho = -0.096$, $p = 0.431$). However, higher TOLT scores were weakly correlated with greater efficiency, indicated by fewer retries ($\rho = -0.259$, $p = 0.032$). Metacognitive analysis revealed high confidence in incorrect answers, suggesting poor calibration. **Conclusions:** Abstract logical reasoning skills do not automatically transfer to novel problem-solving contexts. The pedagogical value of these puzzles lies not in applying existing knowledge but in developing procedural efficiency, metacognitive resilience, and strategic thinking to manage cognitive load.

Keywords: Logical Reasoning; Teacher Training; Skill Transfer; Metacognition; Variant Sudoku.

Corresponding author: Raúl Martínez-Bohórquez. Email: raulmb@unex.es

Proceso sobre Producto: Lógica Formal, Puzles de Sudoku y el Desarrollo del Razonamiento en Docentes en Formación

RESUMO

Contexto: Existe uma desconexão crítica entre o raciocínio lógico formal ensinado a futuros professores e as competências práticas de resolução de problemas necessárias numa sala de aula. Este estudo explora este problema de "transferência distante" utilizando quebra-cabeças lógicos complexos. **Objetivos:** Esta investigação teve como objetivo analisar a relação entre as competências de raciocínio lógico formal de futuros professores, medidas pelo Teste de Pensamento Lógico (TOLT), e o seu desempenho, eficiência e estratégias metacognitivas na resolução de quebra-cabeças de Sudoku com variantes. **Desenho:** O estudo empregou um desenho descritivo-correlacional para analisar a relação entre as competências cognitivas dos participantes e os seus comportamentos na resolução de quebra-cabeças. **Ambiente e Participantes:** O estudo envolveu 69 professores voluntários em formação inicial do curso de Licenciatura em Educação Primária da Universidade da Extremadura, Espanha, durante o ano letivo de 2024/2025. **Coleta e análise de dados:** Os dados foram recolhidos através do TOLT, métricas de resolução de quebra-cabeças (taxas de resolução, tempo, tentativas) e questionários reflexivos. A análise incluiu a correlação de postos de Spearman e a análise temática das respostas qualitativas. **Resultados:** As pontuações no TOLT não previram significativamente o sucesso na resolução dos quebra-cabeças ($\rho = -0.096$, $p = 0.431$). No entanto, pontuações mais altas no TOLT correlacionaram-se fracamente com uma maior eficiência, indicada por um menor número de tentativas ($\rho = -0.259$, $p = 0.032$). A análise metacognitiva revelou uma elevada confiança em respostas incorretas, sugerindo uma fraca calibração. **Conclusões:** As competências de raciocínio lógico abstrato não se transferem automaticamente para contextos novos de resolução de problemas. O valor pedagógico destes quebra-cabeças reside não na aplicação de conhecimentos pré-existentes, mas no desenvolvimento da eficiência procedimental, da resiliência metacognitiva e do pensamento estratégico para gerir a carga cognitiva.

Palavras-chave: Raciocínio Lógico; Formação de Professores; Transferência de Habilidades; Metacognição; Sudoku com Variantes.

INTRODUCTION

The capacity for logical reasoning is a cornerstone of effective teaching (Johnson-Laird, 2012). In an era where 21st-century skills such as critical thinking, adaptive problem-solving, and analytical reasoning are paramount for economic and social progress, primary school teachers are on the front lines of cultivating these abilities in the next generation. Therefore, the development of these cognitive abilities within teacher training programs is of critical importance (Marchis, 2013). Puzzles, with their inherent demand for analytical thought, have long been recognized as valuable educational instruments

(Goldenberg, 2013). Variant Sudokus and other logic puzzles, which introduce additional rules, offer a rich landscape for cognitive engagement (Baek et al., 2008). While previous research has often focused on the use of standard puzzles in K-12 settings, few studies have investigated the use of complex, variant logic puzzles as a tool for developing metacognitive skills in an adult, pre-service teacher population. Teacher education programs thus face a critical challenge: how to move pre-service teachers beyond abstract logical knowledge towards the flexible, resilient, and metacognitive problem-solving required in a real classroom. This study investigates whether complex logic puzzles can serve as a pedagogical tool to bridge this gap.

This paper presents an empirical study exploring the relationship between formal reasoning and applied problem-solving in a novel context. Building on the extensive literature on the challenges of skill transfer—particularly 'far transfer', where abstract knowledge often fails to apply to practical, dissimilar tasks (Perkins & Salomon, 1989)—this study investigates this phenomenon within teacher education. We use complex logic puzzles to explore whether formal reasoning skills, as measured by the Test of Logical Thinking (TOLT; Tobin & Capie, 1981), transfer to the applied problem-solving abilities required in these tasks. The study measured variables such as puzzle resolution rates and retries, alongside self-reported metacognitive states.

The cognitive challenges that make these puzzles interesting for studying human cognition have also led to their adoption in benchmarking Artificial Intelligence (AI). Initiatives like Sakana AI's Sudoku-Bench use such puzzles to test multi-step logical reasoning in advanced AI systems (Seely et al., 2024). This parallel offers an opportunity to tentatively contextualize our findings about human problem-solving against the known challenges these same puzzles pose to AI, enriching our understanding of the puzzles as a tool for studying uniquely human skills.

This paper is structured as follows: First, we review the relevant literature on logical reasoning in teacher education, Cognitive Load Theory, the challenge of skill transfer, and the broader use of puzzles as pedagogical tools. Next, we describe the methodology employed in this study, including details on the participants and instruments. Following this, we present the results of our statistical analysis. Finally, we discuss the implications of our findings for both teacher education and the study of advanced reasoning, alongside the study's limitations and concrete directions for future research.

THEORETICAL BACKGROUND

The ability to think logically is a cornerstone of effective teaching, but its conception in educational research has evolved significantly. Historically, logical thinking was viewed through a Piagetian lens as a general, abstract cognitive skill, with formal reasoning—encompassing abilities like controlling variables, proportional thinking, and combinatorial reasoning—seen as the apex of cognitive development (Piaget & Inhelder, 1958). Standardized instruments like the Test of Logical Thinking (TOLT) were developed from this tradition to provide a quantitative measure of these abilities (Tobin & Capie, 1981).

However, contemporary research frames logical reasoning not as an abstract skill, but as a situated and domain-specific practice. In education, this manifests as an argumentative practice: the ability to construct, communicate, and critically evaluate evidence-based claims (Kuhn, 1999; Osborne, 2010). Within mathematics education, this concept is crystallized in frameworks like Mathematical Knowledge for Teaching (MKT). MKT posits that teachers require a specialized form of reasoning that goes beyond simply solving a problem for themselves. It involves interpreting students' varied and often unconventional thinking, evaluating the logical validity of their strategies, and justifying mathematical procedures in a way that is both conceptually sound and pedagogically accessible (Ball et al., 2008). A teacher's logical competence, therefore, is less about formal symbolic logic and more about this applied pedagogical reasoning.

Despite its criticality, research consistently shows that many pre-service teachers enter their training with still-developing reasoning skills. Foundational studies using the TOLT revealed significant deficits in formal reasoning among prospective teachers (Lawson & Bealer, 1984), and a strong correlation was established between these skills and future academic success in science and mathematics (Bitner, 1991). This challenge directly impacts their future teaching. Empirical studies have shown that pre-service teachers often struggle to evaluate the validity of student arguments, tending to focus on superficial features (e.g., the final answer) rather than the coherence of the underlying logic (Knuth, 2002; Martin & Harel, 1989). This can lead to treating mathematical algorithms as arbitrary sets of rules to be memorized rather than as logical, justifiable processes (Chinnappan & Kajander, 2016).

Crucially, logical thinking does not operate in a vacuum. It is inextricably linked to two other domains. The first is *metacognition*: the capacity to plan, monitor, and evaluate one's own thought processes (Flavell,

1979). For a teacher, this is a dual responsibility; they must be aware of their own reasoning to effectively model and guide their students' reasoning (Aizikovitsh-Udi & Amit, 2011). The second is the *affective domain*, where a student's attitudes, beliefs, and emotions can act as powerful mediators of cognitive performance. Anxiety or a lack of confidence can significantly inhibit the ability to engage in complex reasoning tasks (McLeod, 1992). Recognizing this deep interconnection, integrated models for teacher education now advocate for interventions that simultaneously address cognitive strategies, metacognitive monitoring, and emotional regulation to cultivate resilient and logically competent educators (Blanco et al., 2015). This study is situated within this integrated framework, using logic puzzles as a tool to engage all three domains.

Cognitive load and the architecture of complex problem-solving

To understand why complex puzzles can be effective pedagogical tools, it is essential to consider the architecture of human cognition. *Cognitive Load Theory (CLT)* provides a powerful framework for this, positing that working memory is limited and that instruction should be designed to manage the cognitive demands placed upon it (Sweller et al., 2019). CLT distinguishes between three types of load:

- *Intrinsic Load:* The inherent, unchangeable complexity of the information or task itself. A Killer Sudoku, with its interacting arithmetic and placement rules, has a higher intrinsic load than a classic Sudoku.
- *Extraneous Load:* The unnecessary cognitive work required by the way information is presented. This is often called "bad" load and can be caused by confusing instructions or a poorly designed interface. Effective instruction aims to minimize this.
- *Germane Load:* The "good" load, which refers to the working memory resources devoted to the process of learning—that is, constructing and automating mental schemas.

From a CLT perspective, the logic puzzles used in this study are valuable precisely because they impose a high but manageable *intrinsic* load. By providing clear rules and a familiar grid format, the *extraneous* load is minimized. This frees up the learner's cognitive resources to be invested in *germane* load: the effortful process of identifying underlying logical patterns, developing new problem-solving strategies, and integrating multiple rules. The frustration and effort reported by participants can be reinterpreted as the

experience of high germane cognitive load, a necessary condition for the deep processing required to build robust and flexible problem-solving schemas (Sweller et al., 2019).

The challenge of transfer: from puzzle-solving to pedagogical practice

A central goal of education is the *transfer of learning*: the ability to apply knowledge and skills learned in one context to new and different situations (Perkins & Salomon, 1989). The premise of using puzzles in teacher training rests entirely on this principle—that the logical, metacognitive, and resilient habits of mind developed while solving a puzzle will transfer to the complex, dynamic context of a primary school classroom. However, research has consistently shown that such "far transfer"—from a well-defined puzzle to an ill-defined professional practice—is incredibly difficult to achieve and rarely happens automatically (Barnett & Ceci, 2002).

Successful transfer is not a matter of simple practice; it requires what Perkins and Salomon term "high-road transfer," which is effortful and conscious. It depends on the learner's ability to engage in *mindful abstraction*, decontextualizing a principle from its original learning situation so that it can be applied elsewhere (Perkins & Salomon, 1989). For instance, a student teacher must move from thinking "I am stuck on this Thermo Sudoku, I should re-read the rules and check my assumptions" to the abstract principle "When faced with an unexpected obstacle, I should pause, review the core constraints, and question my initial approach."

This process of mindful abstraction is fundamentally metacognitive. Therefore, a pedagogical intervention aimed at fostering transfer cannot merely provide problems to solve; it must explicitly prompt learners to reflect on their problem-solving process. The reflective questions used in this study—prompting participants to analyse their thoughts and feelings before, during, and after the task—are designed specifically to facilitate this high-road transfer, encouraging them to abstract transferable strategies from the concrete act of puzzle-solving.

Variant sudoku and logic puzzles as cognitive training tools

The use of complex puzzles as instruments for cognitive development extends far beyond mathematics education, forming part of a wider pedagogical movement that leverages well-defined problems to train professional reasoning skills. For instance, legal education has long used logic puzzles and hypothetical scenarios (which function as text-based puzzles) to hone the skills

of deductive and inductive reasoning, forcing students to identify premises, follow logical chains, and spot fallacies. Similarly, in medical curricula, diagnostic puzzles and clinical case challenges are used to develop the hypothetico-deductive reasoning required for accurate diagnosis. In computer science education, puzzles like the Tower of Hanoi are used to teach fundamental concepts of recursion, backtracking, and algorithmic thinking. These examples from other fields strengthen the argument that complex puzzles are not mere pastimes; they are robust simulators for the kind of structured, multi-step reasoning that is a hallmark of professional expertise.

Grounded in this wider context, variant Sudokus and other logic puzzles emerge as powerful instruments for teacher training. While standard Sudoku puzzles are widely known, variants (e.g., Filomino, Tentaisho) offer an expanded range of challenges that demand flexible and creative modes of thinking. Their structured yet varied nature makes them ideal for developing the specific skills essential for teaching. They can help develop:

- *Deductive and Inductive Reasoning:* Applying general rules to specific instances and identifying patterns to infer solutions. This mirrors the pedagogical task of applying a general mathematical concept to a specific student's question.
- *Working Memory and Strategic Thinking:* Holding multiple, interacting constraints in mind simultaneously to plan sequences of moves and anticipate consequences (Baek et al., 2008). This is directly analogous to managing the high cognitive load of a classroom environment.
- *Perseverance and Cognitive Resilience:* Working through complex problems that require backtracking and re-evaluation, thereby strengthening tolerance for ambiguity and initial failure.
- *Metacognitive Monitoring:* The puzzle format provides immediate feedback, forcing the solver to recognize when a chosen path is incorrect and to self-correct. This is a direct, practical application of the metacognitive skills required for mindful abstraction and transfer.

The adaptability of these puzzles, from simpler variants to complex ones with multiple interacting constraints, allows for graduated learning experiences suitable for adult learners, including student teachers. They provide a concrete, engaging context in which to practice the abstract reasoning skills discussed previously.

The human-ai parallel: puzzles as a benchmark for advanced reasoning

The same cognitive demands that make variant Sudokus engaging educational tools also position them as robust benchmarks for artificial intelligence (AI) reasoning. A notable paradox has emerged in AI capabilities: while large language models (LLMs) can achieve high scores on structured assessments like the TOLT, often by leveraging patterns in vast training data, their performance collapses when faced with novel logic puzzles that require creative, multi-step reasoning. This phenomenon, termed the "curse of complexity," highlights a gap between pattern matching and genuine problem-solving, a limitation not easily overcome by simply scaling model size (Lin et al., 2024; Lample & Charton, 2020; Shojaee et al., 2024).

To address this, initiatives like *Sudoku-Bench* by Sakana AI have been developed (Seely et al., 2024). This benchmark, created with puzzle masters from "Cracking The Cryptic" and Nikoli, uses a curated set of challenging and unconventional Sudoku variants to evaluate creative reasoning in AI systems. Its purpose is to move beyond brute-force computation and assess an AI's ability to approximate human-like intuition, meta-reasoning, and the discovery of novel logical steps (Seely et al., 2024). Initial results show that even leading LLMs solve fewer than 15% of these puzzles, confirming their difficulty and their utility in testing the frontiers of long-horizon, strategic reasoning (Seely et al., 2024).

This parallel provides a unique and powerful context for this study. It externally validates that the selected puzzles are not trivial but represent a class of problems known to be challenging for even the most advanced AI systems. By examining how student teachers approach these puzzles—including their logical successes, their management of cognitive load, their metacognitive awareness (or lack thereof), and their affective responses—we can gain insight into the uniquely human aspects of problem-solving that current AI benchmarks are only beginning to explore (Bowman et al., 2021). This allows us to frame the classroom experience not just as a pedagogical exercise, but as an investigation into the nature of complex reasoning itself.

METHODOLOGY

This study employed a descriptive correlational design to investigate the logical reasoning skills, puzzle-solving performance, and metacognitive strategies of future primary school teachers.

Participants

Participants were 69 student teachers from the Degree in Primary Education program at the Universidad de Extremadura, Spain, during the 2024/2025 academic year. Participation was voluntary. Key demographic information is summarized in Table 1.

Written informed consent was obtained from all participants involved in the study. The research project from which this article originates was not submitted for prior evaluation by a research ethics committee, as the activities were considered low-risk and integrated within the standard educational practices of the degree program, with all data being handled anonymously for academic purposes. In accordance with the journal's guidelines, the authors assume full and explicit responsibility for the ethical conduct of the research and hereby exempt *Acta Scientiae* from any consequences arising therefrom. This includes, in accordance with Resolution No. 510 of April 7, 2016, of the National Health Council of Brazil, full responsibility for assistance and eventual compensation for any damages resulting to any of the research participants.

Table 1

Source: Elaborated by the authors (2025).

Characteristic	Value
Gender	46 Female, 23 Male
Age Range	20-25 years
Mean Age (SD)	20.81 (1.15) years

Instruments and materials

Puzzle test instrument

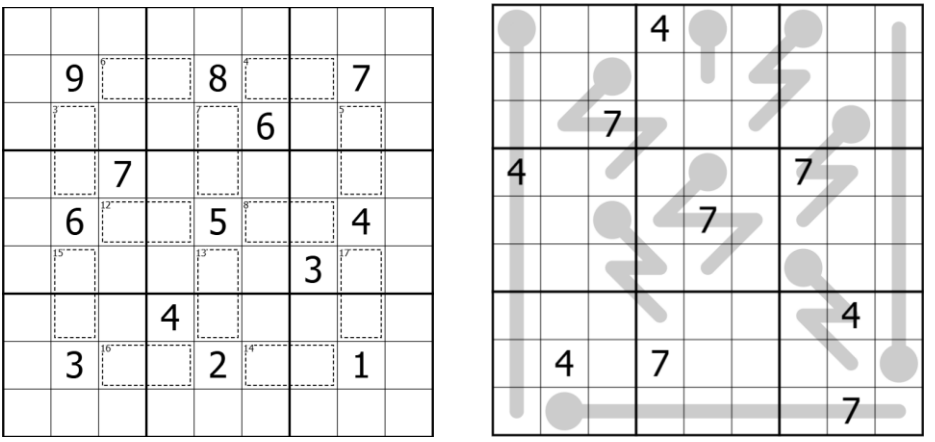
A curated selection of variant Sudoku and other logical pencil puzzles was used for the study. These included:

- Killer sudoku: classical sudoku rules apply (place digits 1-9 in each row, column, and 3x3 box, without repetition), and the sum of the digits inside certain regions (cages) are given as clues.

- Thermo sudoku: classical sudoku rules apply, and digits must strictly increase along given lines (thermometers) from a bulb to the tip.
- Circle sudoku: classical sudoku rules apply, and a digit inside a circle indicates how many circles contain that digit (e.g. a 5 in a circle indicates that exactly 5 circles on the grid contain a 5).
- Fillomino: the grid must be divided into orthogonally connected regions (polyominoes); the number in a cell indicates the size of the polyomino it belongs to, and regions of the same size cannot be orthogonally adjacent.
- Tentaisho: the grid must be divided into orthogonally connected regions, each containing exactly one circle and having 180° rotational symmetry around that circle.

Figure 1

Examples of variant logic puzzles used in the study. (a) A Killer Sudoku, where digits in cages must sum to the indicated total. (b) A Thermo Sudoku, where digits must increase along the marked lines. Source: Puzzles adapted from Logic Masters Deutschland, by Philip Newman and Harry Dorcy, respectively.



Puzzles were sourced from online communities such as Logic Masters Deutschland (logic-masters.de) and Cracking The Cryptic. Instructions for each variant were provided clearly. Puzzles were selected to offer a gradual increase in complexity, starting from the more familiar sudoku variants and up to the more complex pencil puzzles.

For each puzzle, participants were instructed to record specific metrics (problems attempted, problems solved, total resolution time, average time per solution, total retries, and perceived difficulty) and provide a visual representation of their solution. The instrument also included reflective questions about their problem-solving process.

The test of logical thinking (TOLT)

To objectively measure baseline levels in logical reasoning, established psychometric tools are necessary. The Test of Logical Thinking (TOLT), developed by Tobin & Capie (1981), is a widely used instrument for assessing formal operational reasoning. The TOLT evaluates five key modes of formal thought:

- **Controlling Variables (CV):** Identifying and controlling variables in an experimental setup.
- **Proportional Reasoning (PR):** Understanding and applying ratio and proportion concepts.
- **Combinatorial Reasoning (CR):** Systematically considering all possible combinations of a set of items.
- **Probabilistic Reasoning (PbR):** Understanding and applying concepts of chance and probability.
- **Correlational Reasoning (CrR):** Determining the relationship between variables.

The test's robust reliability (Cronbach's alpha typically around .85) and validity make it a suitable instrument for studies examining cognitive development and the impact of specific interventions on logical thinking skills (Tobin & Capie, 1981). Its use in this study allows for a quantitative assessment of the student teachers' logical reasoning abilities.

A Spanish version of the test was administered to the students. This instrument consists of 10 items. For each item, participants provided an answer and a justification. A point was awarded only when both were correct, leading to a total score out of 10. Additionally, for each item, participants rated their confidence in their answer on a 5-point Likert scale (1=not sure at all, 5=completely sure).

Questionnaires

- **Demographic questionnaire:** Collected basic information (age, gender, prior puzzle-solving experience). This was administered prior to the main sessions.
- **Reflective questions (integrated into puzzle test):** As part of the puzzle test instrument, participants responded to open-ended questions regarding their thoughts and feelings based on the phases of the Integrated Model of Mathematics Problem Solving (IMMPS), which integrates cognitive and affective-metacognitive components (Blanco et al., 2013; Blanco et al., 2015).
    - Before starting a mathematical problem (What do I think? What do I tell myself? How do I feel? What do I do as a consequence of these thoughts and feelings?).
    - While trying to solve a mathematical problem (What do I think? What do I tell myself? How do I feel? What do I do as a consequence of these thoughts and feelings?).
    - After solving a mathematical problem (What do I think? What do I tell myself? How do I feel? What do I do as a consequence of these thoughts and feelings?).

Procedure

The study was conducted in two distinct sessions, each lasting approximately 1 hour and 30 minutes. Prior to these sessions, participants provided informed consent and completed a demographic questionnaire.

1. **Session 1: Puzzle Test Administration (1.5 hours)**
    - Participants were provided with the puzzle test instrument.
    - Instructions were given to solve each of four puzzles (the Killer sudoku, the Tentaisho, the Fillomino and either the Thermo or the Circle sudokus) by applying the Integrated Model of Mathematics Problem Solving (IMMPS), as oriented in their class.
    - For each puzzle, participants were required to record:
        - The time taken to place the first digit/make the first significant move they felt certain about.

        - The number of retries or significant backtracks.
        - The total time employed to solve the puzzle.
        - An estimated difficulty score on a scale of 0 (very easy) to 10 (very difficult).
        - A visual representation (e.g., a screenshot) of the completed puzzle solution.
    - After attempting the puzzles, participants responded to the sets of reflective questions concerning their thoughts, self-talk, feelings, and consequent actions before, during, and after the problem-solving process.

2. **Session 2: Test of Logical Thinking (TOLT) Administration (1.5 hours)**
    - The Spanish-translated version of the Test of Logical Thinking (TOLT) was administered to the participants.
    - Standardized instructions for the TOLT were provided. Participants completed the test individually within the allocated session time.

Data analysis

Data from the TOLT and puzzle-solving instruments were merged using the Python library Pandas and analysed with RStudio. The analysis included:

- **Descriptive Statistics:** Mean, median, standard deviation (SD), and frequency distributions were calculated for TOLT scores, puzzle performance metrics (resolution rate, time, retries, perceived difficulty), and students' self-reported confidence levels.
- **Metacognitive Calibration Analysis:** The average confidence level (on a 1-5 scale) was compared between correctly answered and incorrectly answered items on the TOLT to assess students' self-monitoring accuracy.
- **Thematic Analysis:** Qualitative responses regarding participants' thoughts and feelings before, during, and after problem-solving were categorized as positive, negative, or neutral to identify dominant metacognitive and affective patterns at each stage.

- **Correlational Analysis:** Spearman's rank correlation coefficient was used to explore the relationships between TOLT scores and key puzzle performance metrics (e.g., resolution rate, retries) due to the non-parametric nature of some data.

RESULTS AND ANALYSES

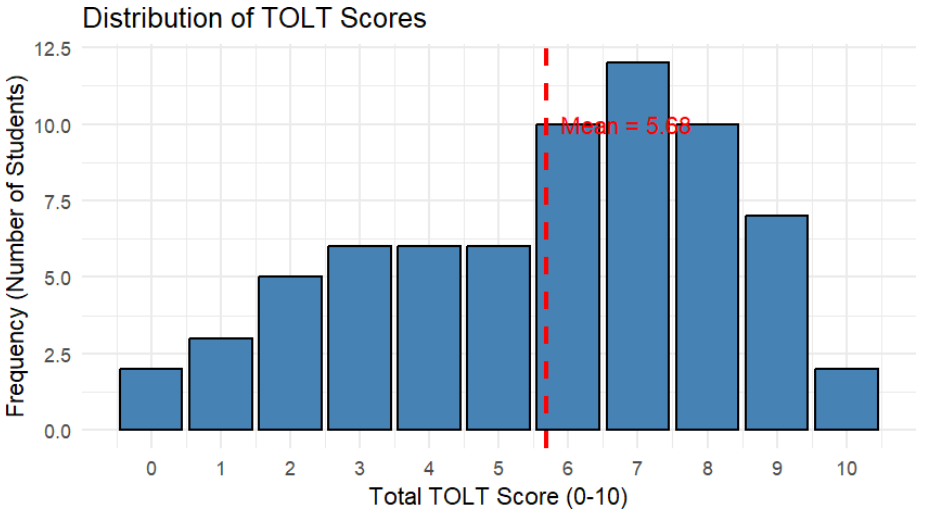
The analysis of the merged dataset from 69 student teachers yielded key insights into their logical thinking abilities, puzzle-solving performance, and the interplay with their metacognitive strategies.

Logical thinking profile and metacognitive calibration

The mean total score on the TOLT was 5.68 (SD = 2.58) out of 10, indicating a developing level of formal logical thinking with significant heterogeneity (Figure 2).

Figure 2

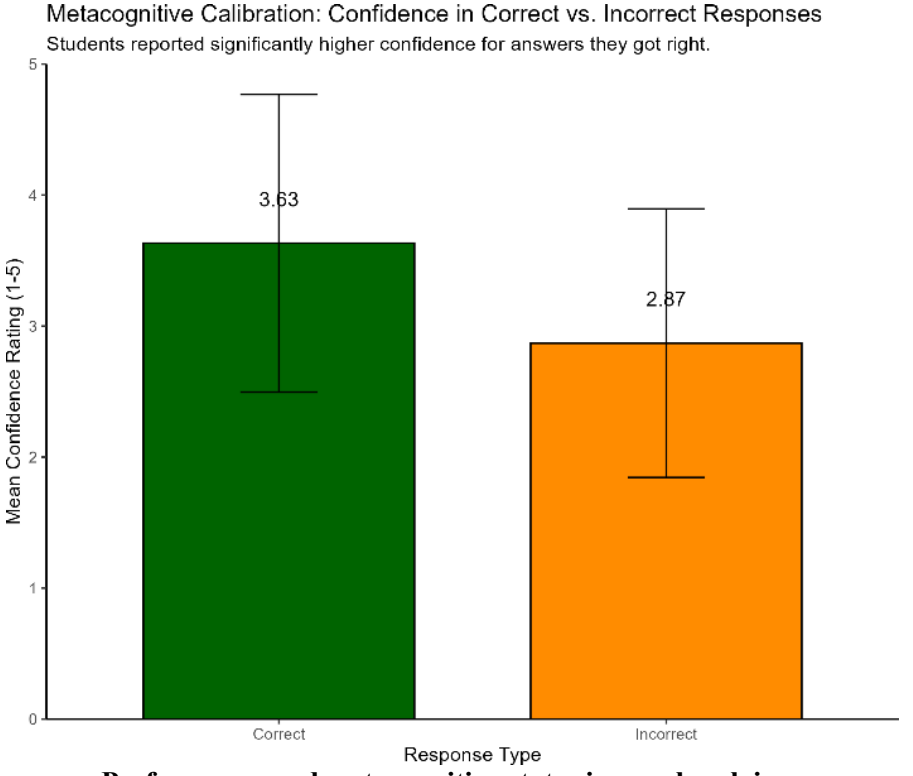
Metacognitive calibration analysis (Figure 3) shows that while participants were more confident in their correct answers (M = 3.63) than their incorrect ones (M = 2.87), the confidence level for incorrect answers was notably high.This suggests a degree of poor metacognitive calibration, or an "illusion of knowing."



Distribution of TOLT scores among Student Teachers. The mean score was 5.68. Source: Elaborated by the authors (2025).

Figure 3

Metacognitive Calibration on the TOLT. This bar chart displays the mean self-reported confidence (on a 1-5 scale) for correct versus incorrect answers. The high confidence level for incorrect answers (M = 2.87) indicates poor metacognitive calibration. Source: Elaborated by the authors (2025).

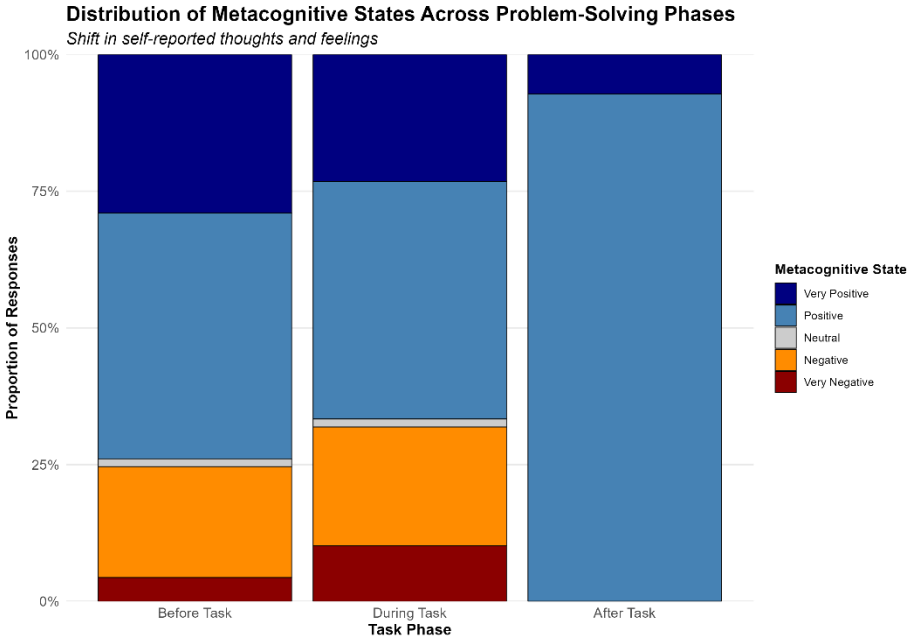


Performance and metacognitive states in puzzle solving

Participants demonstrated high engagement, with an average resolution rate of 85.9% (SD = 21.6). On average, they spent 84.6 minutes (SD = 45.4) and required 7.94 retries (SD = 6.90). Thematic analysis of reflective questionnaires (Figure 4) revealed a distinct affective journey: a majority felt positive before the task (73%), which dropped slightly during the task with an increase in negative reports (frustration), followed by a dramatic shift to overwhelmingly positive feelings upon completion.

Figure 4

Distribution of Metacognitive States Across Problem-Solving Phases. Source: Elaborated by the authors (2025).



Deepening the analysis: the disconnect between formal logic and puzzle performance

Correlational analysis revealed a nuanced relationship between formal logical thinking and practical problem-solving. Consistent with the 'far transfer' problem, Spearman's correlation showed *no significant relationship* between the total TOLT score and the puzzle resolution rate ($\rho = -0.096$, $p = 0.431$). This indicates that a student's general level of formal logical reasoning did not predict their ultimate success in solving the puzzles.

However, a weak but statistically significant negative correlation was found between the total TOLT score and the number of retries ($\rho = -0.259$, $p = 0.032$). To investigate this further, we disaggregated the TOLT score into its five constituent sub-scales (descriptive statistics in Table 2).

Table 2

Descriptive Statistics for TOLT Sub-scale Scores (Max Score = 2). Source: Elaborated by the authors (2025).

TOLT Sub-scale	Mean	Standard Deviation (SD)
Controlling Variables (CV)	1.29	0.79
Proportional Reasoning (PR)	0.88	0.88
Combinatorial Reasoning (CR)	1.10	0.86
Probabilistic Reasoning (PbR)	0.99	0.81
Correlational Reasoning (CrR)	1.42	0.81

The subsequent correlational analysis (Table 3) robustly confirmed the primary finding, as no statistically significant correlation was found between any of the five TOLT sub-scales and the puzzle resolution rate. Notably, even Combinatorial Reasoning (CR), theoretically aligned with Sudoku, showed no significant relationship with success ($\rho = -0.209$, $p = 0.085$).

This granular analysis did not yield a statistically significant predictor for process efficiency among the sub-scales. It is noteworthy, however, that the correlation between the Controlling Variables (CV) score and a lower number of retries approached the significance threshold ($\rho = -0.233$, $p = 0.054$). While this result is not statistically significant by the $p < .05$ standard, it may suggest a potential avenue for future research with larger samples.

Table 3

Spearman's Rho (ρ) and p-values for TOLT Sub-scales vs. Puzzle Performance Metrics. Source: Elaborated by the authors (2025).

TOLT Sub-scale	Resolution Rate	Total Time	Retries
Controlling Variables	-0.162 (p=.182)	-0.116 (p=.343)	-0.233 (p=.054)
Proportional Reasoning	-0.142 (p=.243)	0.040 (p=.746)	-0.155 (p=.205)
Combinatorial Reasoning	-0.209 (p=.085)	-0.013 (p=.916)	-0.172 (p=.159)
Probabilistic Reasoning	0.061 (p=.621)	0.131 (p=.283)	-0.101 (p=.408)
Correlational Reasoning	0.082 (p=.503)	0.036 (p=.768)	-0.091 (p=.455)

DISCUSSION

This study was designed to explore the interplay between formal logical reasoning, metacognition, and performance on complex logic puzzles among future teachers. The findings, however, reveal a compelling and counter-intuitive disconnect: a student's level of formal logical reasoning, as measured by the TOLT, did not predict their success in solving the puzzles. This lack of a significant correlation ($\rho = -0.096$) between TOLT scores and puzzle resolution rates provides a powerful empirical illustration of the well-documented challenge of learning transfer.

The "transfer problem" in action

The central premise of many educational interventions is that abstract skills will be applied in new contexts. Our findings challenge this assumption directly. The skills assessed by the TOLT—controlling variables, proportional reasoning, etc.—are foundational to scientific and mathematical thought. Yet, they did not confer a direct advantage in solving these novel logic puzzles. This supports the argument by Perkins & Salomon (1989) that such "far transfer" is not automatic and requires a conscious, effortful process of mindful abstraction. The student teachers did not, or could not, spontaneously map their abstract logical knowledge onto the specific, practical demands of the puzzles.

This result reframes the pedagogical value of these puzzles. They are not merely a field for *applying* pre-existing logical skills. Instead, they represent a distinct cognitive domain that requires its own set of strategies. The puzzles demand not just logic, but also cognitive flexibility, strategic thinking in the face of uncertainty, and resilience—skills that the TOLT is not designed to measure.

Notably, even the TOLT sub-scale for Combinatorial Reasoning, which theoretically aligns most closely with the demands of a Sudoku-style puzzle, failed to predict success. This underscores the profound gap between possessing a decontextualized logical skill and applying it effectively within a complex, constrained problem space.

Process over outcome: interpreting the role of logic

While formal logic did not predict the final outcome (success), it did appear to influence the *process*. The weak but significant negative correlation between TOLT scores and the number of retries ($\rho = -0.259$) suggests that students with stronger formal reasoning were more efficient. They made fewer significant errors that required starting over. The near-significant correlation

between the "Controlling Variables" sub-scale and retries ($\rho = -0.233$, $p = .054$) offers a potential explanation: students adept at systematically isolating and testing variables may be better at methodically checking their work and catching errors early, even if this doesn't guarantee they will find the key insight needed to solve the entire puzzle.

This finding aligns with Cognitive Load Theory (Sweller et al., 2019). A student with a more automated schema for logical checking (e.g., controlling variables) may reduce their own extraneous cognitive load, freeing up working memory to grapple with the puzzle's high intrinsic load. Their problem-solving process is less chaotic, even if their ultimate success depends on other factors like creativity or a flash of insight.

Metacognition and the affective journey

The disconnect between formal logic and success places even greater weight on the roles of metacognition and affect. The participants' journey—from pre-task positivity, through during-task frustration, to post-task satisfaction—is the story of grappling with a challenge that their existing knowledge could not easily solve. The high confidence placed in incorrect answers ($M=2.87$) further highlights this struggle; in the absence of a clear, transferable algorithm, students were forced to rely on heuristics and intuition that were sometimes flawed, yet they remained overly confident.

This underscores the need for teacher education to focus on building metacognitive awareness for navigating ambiguity. The goal should not be to just equip students with logical tools, but to teach them how to react when those tools are not enough. The overwhelmingly positive feeling after completion suggests that successfully navigating this struggle is intrinsically rewarding and may build the resilience needed for future, ill-defined problems they will face in the classroom.

The pedagogical value of complex puzzles: insights from the human-AI divide

The findings of this study gain a deeper significance when contextualized against the landscape of AI. A notable paradox in current AI capabilities is the phenomenon termed the “curse of complexity,” where models excel on structured assessments but fail when faced with novel, complex logic puzzles that require creative, multi-step reasoning (Lin et al., 2024; Lample & Charton, 2020; Shojaee et al., 2024). Standardized tests often fall short in assessing this type of creative problem-solving. Recent work has shown that even advanced reasoning models experience a "complete accuracy collapse" on

puzzles like the Tower of Hanoi once a certain complexity threshold is reached, sometimes even when provided with the correct algorithm (Shojaee et al., 2024).

Our study's empirical human data provides a compelling parallel to this. Just as AI models struggle to transfer generalized knowledge to specific, novel tasks, we found that proficiency in a formal logic test does not directly translate to success in an applied reasoning task. The challenges our student teachers faced—the emotional journey, the metacognitive failures, and the disconnect between formal knowledge and practical application—provide a rich, human-centric dataset that mirrors these frontiers in AI. The unique constraints of variant Sudoku puzzles, which challenge AI to identify novel logical breakthroughs rather than relying on pre-learned patterns, are a key tool in this investigation. Initiatives like Sakana AI's Sudoku-Bench are designed specifically to evaluate this creative, multi-step logical reasoning, with initial results showing that leading AI models solve fewer than 15% of these puzzles (Seely et al., 2024).

This suggests that a robust evaluation of advanced reasoning, in both humans and AI, must encompass more than solution accuracy. Pedagogically, these puzzles should not be used as summative assessments of logic, but as formative tools within a 'cognitive apprenticeship' model (Collins et al., 1991). In such a model, instructors can make their own thinking visible by modeling solving processes, coaching students through difficulties, and encouraging articulation and reflection through 'think-aloud' protocols. For instance, an instructor could solve a challenging Thermo Sudoku for the class, verbalizing not just the logical deductions, but also their moments of uncertainty, their reasons for abandoning a failing strategy, and their metacognitive checks. This approach shifts the focus from finding the answer to developing the metacognitive and strategic skills essential for teaching. The "overconfidence-in-error" phenomenon observed in our study represents a specific, measurable human cognitive bias that current AI systems typically lack. While our findings are preliminary, they point toward the potential value of incorporating metrics of process, efficiency (e.g., retries), and metacognition into the design of more holistic AI benchmarks (Bowman et al., 2021).

Limitations and future research directions

While this study provides insights, its limitations must be acknowledged. As an exploratory work, the primary limitation is the **cross-sectional and correlational design**, which precludes causal inference. Further limitations include:

- *Statistical Power and Sample Size:* With a modest sample size of N=69, this study was likely underpowered to detect small to medium effect sizes. Consequently, the non-significant findings (e.g., the lack of correlation between TOLT scores and success) should not be interpreted as definitive proof of an absence of a relationship, but rather as a failure to detect one with the current sample.
- *Instrument Validity:* While the TOLT is a classic instrument, future research might consider more contemporary measures of reasoning that align with the multifaceted skills required in modern teaching contexts, such as argumentative writing or scenario-based assessments.
- *Nature of the Puzzles:* The selection of puzzles, while varied, represents a specific genre of well-defined logic problems. The findings might not generalize to other forms of problem-solving, such as ill-structured problems or the complex social reasoning required in a classroom.
- *Absence of a Control Group:* The study lacks a control group, making it difficult to isolate the specific effects of the logic puzzles.
- *Reliance on Self-Report Data:* The assessment of metacognitive states relied on self-reports, which can be subject to bias.
- *Sample Generalizability:* The findings may not be generalizable beyond the specific context of this study.

These limitations inform avenues for future research, such as longitudinal designs, the use of active control groups, and mixed-methods approaches (e.g., think-aloud protocols) to capture richer data on problem-solving strategies.

CONCLUSION

This study set out to investigate the link between formal logic and puzzle-solving in student teachers, but its most significant contribution lies in the absence of that link. We found no significant correlation between scores on the Test of Logical Thinking and the ability to solve complex logic puzzles, providing a stark illustration of the "transfer problem" in a practical, cognitive context. This demonstrates that abstract reasoning skills are not a guarantee of success in novel problem-solving domains.

However, the research also reveals a more nuanced picture. While formal logic did not predict success, it did correlate with a more efficient

problem-solving process, suggesting that such skills help manage cognitive load and reduce errors. The true determinants of success in these tasks may lie in factors the TOLT does not measure: cognitive flexibility, strategic resilience, and creative insight. The participants' affective journey from frustration to satisfaction highlights that the primary pedagogical value of these puzzles may be in training students to manage the uncertainty and high cognitive load inherent in problems for which they have no ready-made algorithm.

The implication for teacher education is clear: it is not enough to teach logical principles in the abstract. We must create learning environments that force students to grapple with the challenge of applying knowledge in new and unexpected ways. By pairing such tasks with explicit metacognitive reflection, we can help future teachers move beyond merely "knowing" logic to developing the wisdom of how and when to use it—and what to do when it is not enough.

AUTHORS' CONTRIBUTIONS STATEMENTS

RMB: Conceptualization, Investigation, Data Curation, Writing – Original Draft. LM: Conceptualization, Investigation, Data Curation, Writing – Review & Editing. LBB: Formal Analysis, Software, Visualization. All authors have read and agreed to the published version of the manuscript

DATA AVAILABILITY STATEMENT

The data presented in this study are available on request from the corresponding author.

FUNDING

This research received no external funding.

INFORMED CONSENT STATEMENT

Informed consent was obtained from all subjects involved in the study.

ACKNOWLEDGEMENTS

We thank the student teachers who participated in this study. We also acknowledge the pioneering work of Sakana AI on Sudoku-Bench (Seely et al., 2024) and the contributions of puzzle creators and communities like "Cracking The Cryptic" for popularizing these engaging variant Sudoku puzzles. We also acknowledge Tobin and Capie for the development of the Test of Logical Thinking.

This work was supported by the Universidad de Extremadura through the Grants for Teaching Innovation Projects (Reference: 2025/2026-LP1-4) for the academic year 2025/2026, within the framework of the Teaching Innovation Group COGNET.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

REFERENCES

Aizikovitsh-Udi, E., & Amit, M. (2011). Developing the skills of critical and creative thinking by probability teaching. *Procedia-Social and Behavioral Sciences, 15*, 1087–1091. https://doi.org/10.1016/j.sbspro.2011.03.243

Baek, Y., Kim, B., Yun, S., & Cheong, D. (2008). Effects of two types of Sudoku puzzles on students' logical thinking. In M. Stansfield & T. Connolly (Eds.), *Proceedings of the 2nd European Conference on Games Based Learning* (pp. 19-24). Academic Conferences and Publishing International Limited.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Bitner, B. L. (1991). Formal operational reasoning modes: Predictors of critical thinking abilities and grades assigned by teachers in science and mathematics for students in grades nine through twelve. *Journal of Research in Science Teaching, 28*(3), 265–274.

Blanco, L. J., Cárdenas, J. A., & Caballero, A. (Eds.). (2015). *La resolución de problemas de matemáticas en la formación inicial de profesores de primaria*. Servicio de Publicaciones, Universidad de Extremadura.

Blanco, L. J., Guerrero, E., & Caballero, A. (2013). Cognition and affect in mathematics problem solving with prospective teachers. In *The Mathematics Enthusiast, 10*, 335–358. University of Montana.

Bowman, S. R., Dror, R., Rogers, A., & Potts, C. (2021). *What will it take to fix benchmarking in natural language understanding?* [Preprint]. arXiv. https://arxiv.org/abs/2104.02145

Chinnappan, M., & Kajander, A. (2016). Pre-service teachers' mathematical understanding: Searching for differences based on school curriculum background. *Fields Mathematics Education Journal, 1*, 3–20.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator, 15*(3), 6–11, 38–46.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911.

Goldenberg, E. P. (2013). The aesthetics of mathematical experience. *For the Learning of Mathematics, 33*(2), 20–25.

Johnson-Laird, P. N. (2012). The history of the study of reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 3–24). Oxford University Press.

Knuth, E. J. (2002). Teachers' conceptions of proof in the context of secondary school mathematics. *Acta Didactica Universitatis Comenianae, Mathematics, 5*, 1–20.

Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16–46.

Lample, G., & Charton, F. (2020). Deep learning for symbolic mathematics. In *International Conference on Learning Representations*. https://arxiv.org/abs/1912.01412

Lawson, A. E., & Bealer, J. M. (1984). The acquisition of basic quantitative reasoning skills during adolescence: Learning or development? *Journal of Research in Science Teaching, 21*(5), 417–423.

Lin, B. Y., Le Bras, R., Richardson, K., Sabharwal, A., Poovendran, R., Clark, P., & Choi, Y. (2024). *ZebraLogic: On the scaling limits of LLMs for logical reasoning* [Preprint]. arXiv. https://arxiv.org/abs/2402.01100

Marchis, I. (2013). Pre-service primary school teachers' logical reasoning skills. *Acta Didactica Napocensia, 6*(4), 59-66.

Martin, W. G., & Harel, G. (1989). Proof frames of preservice elementary teachers. *Journal for Research in Mathematics Education, 20*(1), 41–51.

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 575–596). Macmillan.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*(5977), 463–466.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*(1), 16–25.

Piaget, J., & Inhelder, B. (1958). *The growth of logical thinking from childhood to adolescence*. Basic Books.

Seely, J., Imajuku, Y., Zhao, T., Cetin, E., & Jones, L. (2024). *Sudoku-Bench: Evaluating creative reasoning with Sudoku variants* [Preprint]. arXiv. https://arxiv.org/abs/2405.13600

Shojaee, P., Jones, L., Seely, J., Imajuku, Y., Zhao, T., Cetin, E., Bieber, D., & Samar, A. (2024). *The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity* [Preprint]. arXiv. https://arxiv.org/abs/2406.06941

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 25 years later. *Educational Psychology Review, 31*, 261–292.

Tobin, K., & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement, 41*(2), 413–423.