Higher-Level Domain-General Skills in Math Problem Solving

David W. Braithwaite1 and Jeff Shrager2

1 Department of Psychology, Florida State University, braithwaite@psy.fsu.edu. (Corresponding author)

2 Bennu Climate and Symbolic Systems Program, Stanford University (Adjunct), jshrager@gmail.com.

# Abstract

 Mathematical problem solving relies in part on higher-level domain-general skills, but contributions of such skills to problem solving are understudied in the field of math cognition. This study sought evidence for domain-general contributions to math problem solving of four such skills: logical reasoning, backward reference, planning/evaluating, and allocation of time. Participants completed problem solving tasks in two domains, probability and geometry. Concurrent engagement of each skill during each task was assessed. Individual differences in engagement of logical reasoning were correlated between tasks and associated with accuracy on each task; this pattern of results was not found for the other three skills. Qualitative analysis revealed that errors involving valid logic applied to false premises occurred in both tasks, but errors involving invalid logic occurred primarily on the geometry task. The findings are consistent with logical reasoning contributing to math problem solving, but suggest that the mechanisms underlying such contributions may vary between mathematical domains.

**Keywords:** mathematical cognition, problem solving, logical reasoning, individual differences

# Higher-Level Domain-General Skills in Math Problem Solving

Problem solving in math depends not only on knowledge of particular content domains—such as arithmetic, algebra, probability, or geometry—but also on domain-general skills. These include lower-level cognitive capacities such as working memory (Peng et al., 2016) and spatial thinking (Hawes et al., 2022), and higher-level skills such as logic and planning. The importance of such higher-level skills has long been recognized in the psychology of reasoning (e.g., Ackerman & Thompson, 2017; Johnson-Laird, 1999) and in math education (e.g., Muncer et al., 2022; Schoenfeld, 1992). However, such skills have been understudied in math cognition, which has focused more on domain-specific knowledge and lower-level domain-general skills.

 The present study investigated four higher-level skills that plausibly might contribute to math problem solving in a domain-general way: logical reasoning, backward reference, planning/evaluating, and allocation of time. We assessed individual differences in engagement of each skill while performing problem solving tasks in probability and geometry. We reasoned that, if a given skill contributes to performance on a given task, then individual differences in engagement of that skill during that task should be correlated with accuracy on the task. Further, if the contribution of the skill is domain-general, then engagement of the skill should be correlated between the two tasks. We tested these predictions for each of the four skills above.

 Probability and geometry were selected as the domains of inquiry for three reasons. First, validated problem solving tasks that seemed likely to elicit the use of higher-level skills were available in both domains—for probability, the Probabilistic Reasoning Scale (PRS; Primi et al., 2017), and for geometry, the Geometric Proof Justification Task (GPJT; Braithwaite, 2022). Second, prior research found relations between these tasks at the level of individual differences, which could reflect contributions of domain-general skills to performance on both tasks (Braithwaite, 2022). Third, the PRS and GPJT involve very different content, rendering it unlikely that relations between them reflect shared reliance on particular content knowledge.

 First, we discuss the mathematical domains and tasks that were used in the present study. Then, we define the four candidate skills mentioned above and discuss their potential contributions to problem solving. Finally, we describe the present study in more detail.

## Mathematical Domains and Tasks for Studying Higher-Level Domain-General Skills

### Probabilistic Reasoning and the Probabilistic Reasoning Scale

 Probabilistic reasoning is reasoning about the probabilities of events—for example, the probability that drawing a card from a deck will yield a face card. Probabilistic reasoning is a multifaceted skill that includes the abilities to interpret various representations (e.g., decimals, fractions, percentages) of probability and proportion, to reason about frequency information presented in different formats (e.g., text, tables), to calculate and estimate simple, marginal, joint, and conditional probabilities, and to resist common biases and fallacies (Primi et al., 2017). Individuals vary considerably in probabilistic reasoning skill (Cokely et al., 2012). As these individual differences are related to differences in decision making (Sobkow et al., 2020), differences in probabilistic reasoning may impact outcomes in areas of life that involve probability, such as health and personal finance.

 The PRS is a measure of probabilistic reasoning skill that was developed by Primi and colleagues (2017). It consists of multiple choice probability word problems that assess all aspects of probabilistic reasoning mentioned above. Some items can be solved by calculation—for example, “60% of the population in a city are men and 40% are women. 50% of the men and 30% of the women smoke. We select a person from the city at random. What is the probability that this person is a smoker? (a) 42%, (b) 50%, (c) 85%.” However, items of this type also require reasoning because it is not obvious which calculations are needed to solve them. The PRS also includes items that can be solved by reasoning without calculation—for example, “A bingo game is played with 25 numbers (from 1 to 25). At the first draw, which of the following results is the most likely? (a) It is more likely to be an even number, (b) It is more likely to be an odd number, (c) It is just as likely to be an even or an odd number.”

### Geometric Proof and the Geometric Proof Justification Task

 Proof has been referred to as “the soul of mathematics” (Schoenfeld, 2009). It serves multiple functions in math, including verification, explanation, discovery, and communication (Stylianides et al., 2017). Among mathematicians, deductive logic is central to proof, but many students have difficulty adopting this perspective, instead adopting understandings of proof in which authority or empirical evidence play a greater role (Harel & Sowder, 2007). The development of proof understanding involves a progression from naïve, example-based thinking to grasping abstract deductive structure (Ahmadpour et al., 2019). Further, students often evaluate proofs based on surface features, overlook missing warrants, and otherwise apply criteria different from those of expert mathematicians (Sommerhoff & Ufer, 2019).

In the present study, we focus on geometric proof, which typically constitutes most US students’ first and only exposure to formal mathematical proof (Herbst et al., 2017). Geometric proof depends on many skills that are specifically related to geometry or spatial thinking, such as measurement (Szilágyi et al., 2013), visualization (Battista, 1990), construction and manipulation of diagrams (Cirillo & Hummer, 2021; Clements et al., 2004), and knowledge of geometric terms and definitions (Sinclair & Moss, 2012; Yang & Lin, 2008). However, Braithwaite (2022) argued that some aspects of geometric proof involve skills that are not specific to geometry or spatial thinking, and that one of these aspects is justification—that is, identifying reasons (e.g., principles, definitions, or theorems) why mathematical statements are true.

 To study justification in the context of geometric proof, Braithwaite (2022) created the GPJT (Figure 1). In it, participants are presented with a geometric diagram accompanied by several “given” statements and a “to prove” statement, followed by a proof in two-column format. Each step of the proof consists of a claim, which is provided to the participant, and space for a justification, which is initially blank. The participant’s task is to choose, from a drop-down list, an appropriate justification for each claim. Doing so correctly requires not only geometric knowledge, but also reasoning about logical relations among statements—a skill which is also plausibly involved in many other math problem solving tasks. Figure 1 shows what one proof in the GPJT would look like after the correct justification was chosen for each step.

==================== Insert Figure 1 about here ====================

### Relations Between Probabilistic Reasoning and Geometric Proof Justification

 Braithwaite (2022) found that among university students, accuracies on the PRS and GPJT were correlated, consistent with the tasks relying on overlapping competencies. Correlations among mathematical tasks are often attributed to shared reliance on numerical abilities (Siegler et al., 2011), but the above relation could not be explained in this way because the GPJT does not involve numbers. This relation remained when controlling for algebra skill, suggesting that it did not entirely reflect general math ability or lower-level capacities that are correlated with general math ability, such as working memory. These findings merit further investigation, not only because of the intrinsic importance of probabilistic reasoning and geometric proof, but also because relations between them might provide insight into how higher-level skills contribute to math problem solving across different domains of math.

## Candidates for Higher-Level Domain-General Skills in Math Problem Solving

### Logical Reasoning

 In logic, an argument is said to be valid if and only if its conclusion necessarily follows from its premises—that is, if it is impossible that the premises are true and the conclusion false. We use “logical reasoning” to refer to reasoning that aims to generate valid arguments or to determine whether arguments are valid. Humans routinely commit errors in such reasoning, generating or accepting invalid arguments and rejecting valid ones (Khemlani & Johnson-Laird, 2012; Oberauer, 2006; Wason, 1968). Further, individuals differ substantially in when and how often they commit such errors (Evans et al., 2007; Thompson & Markovits, 2021).

 Logical reasoning is central to mathematics. While other higher-level cognitive skills, such as analogy (Novick & Holyoak, 1991), are important for exploration and discovery in math, only logical reasoning can establish with certainty the truth of mathematical statements[[1]](#footnote-2) (Wu, 1996). Reflecting this critical role of logical reasoning, fostering logical reasoning is an important goal of math education. The Common Core State Standards for Mathematics indicate that students should learn to “build a logical progression of statements to explore the truth of their conjectures” and to “distinguish correct logic or reasoning from that which is flawed” (CCSSI, 2010, Math Practice Standard 3).

 Individuals who have received more mathematical training tend to be better at logical reasoning (Attridge et al., 2015; Attridge & Inglis, 2013; Inglis & Simpson, 2009). This fact is consistent with the possibility that math education improves logical reasoning skills. However, logical reasoning may also contribute to other math skills. Indeed, individual differences in logical reasoning are associated with differences in math problem solving among both children (Battista, 1990; Kleemans et al., 2018; Nunes et al., 2007; Wong, 2018) and adults (Braithwaite, 2022; Morsanyi et al., 2018). Further, training in deductive logic has been found to improve math achievement (Nunes et al., 2007).

 Modern theories of logical reasoning agree that it depends in part on domain-specific knowledge (Khemlani & Johnson-Laird, 2022; Oaksford & Chater, 2020). However, evidence suggests that individual differences in logical reasoning are not entirely domain-specific; instead, they reflect both domain-general and domain-specific contributions. For example, Braithwaite and Rafferty (under review) recently found that both algebra knowledge and logical reasoning with abstract content independently predicted university students’ logical reasoning about algebra. More directly relevant to the present study, logical reasoning with abstract content has been found to predict accuracy on both the PRS (Braithwaite, 2022; Morsanyi et al., 2018) and the GPJT (Braithwaite, 2022), even when controlling for various domain-specific skills. Further, in Braithwaite (2022), abstract logical reasoning also partially mediated the relation between accuracies on the PRS and GPJT, suggesting that shared reliance on the domain-general component of logical reasoning might partially explain that relation. The present study sought more direct evidence for this hypothesis.

### Backward Reference

 We use the phrase “backward reference” to describe returning, after beginning a problem, to the problem statement or earlier steps of one’s work on the problem. Our interest in this behavior was motivated by findings from Braithwaite (2022) with the GPJT. In that task, some trials require justifying claims by stating that they were “given.” One can answer these trials correctly by noticing that the claims appear in the “given” part of the problem statement. Yet, participants erred on 24% of these trials, and accuracy on these trials uniquely predicted accuracy on the PRS when controlling for logical reasoning. Further, some individuals appeared to treat each step of the GPJT in isolation, ignoring what had been established previously. For example, when working on the fourth step of the proof in Figure 1, they ignored the first three steps of the proof, although those steps were necessary for the fourth step to make sense.

 To explain these findings, we reasoned that some individuals may solve problems in a “forward, always” mode, relying on the problem statement only when beginning a problem, and proceeding as directly as possible to an answer. Individuals who, to the contrary, engage in backward reference might be more able to to elaborate implications of given information and integrate old and new information into a coherent whole. Consistent with this possibility, individual differences in understanding or elaboration of given information have been linked to differences in problem solving proficiency in geometry (Cirillo & Hummer, 2021), arithmetic (Cummins et al., 1988; Hegarty et al., 1995), and algebra (Nathan et al., 1992; Sebrechts et al., 1996). Further, training students to elaborate given information before solving problems improves accuracy (Glenberg et al., 2012). However, we are not aware of direct evidence that individuals differ in their tendency to engage in backward reference or that such differences are associated with problem solving accuracy. The present study tested these hypotheses.

### Planning and Evaluating

 An influential theoretical framework in education research (Garofalo & Lester, 1985) highlighted the importance to math problem solving of not only cognitive but also metacognitive skills, which monitor and regulate cognitive ones. There is substantial evidence that metacognitive skills contribute to math problem solving competence, above and beyond the contributions of cognitive skills. Specifically, math problem solving performance is predicted by differences in metacognitive skill (Schoenfeld, 1992; Stillman & Galbraith, 1998; Veenman & van Cleef, 2019), and is improved by interventions that target metacognitive skills in math (Cornoldi et al., 2015; Desoete et al., 2003; Hacker et al., 2019; Mevarech & Kramarski, 2003).

 A long standing but still unresolved issue is whether metacognitive skills are domain-general (Zhao et al., 2019). The present study investigated this possibility, focusing on two metacognitive skills—planning and evaluating. Planning refers to reasoning about what steps to take in solving a given problem (Newell & Simon, 1972); doing so could contribute to probabilistic reasoning by helping individuals to avoid intuitive but incorrect responses (Stanovich et al., 2016; Tversky & Kahneman, 1993), and to performance in geometric proof by promoting understanding of relations among proof steps (Koedinger & Anderson, 1990; Yang, 2012). Evaluating refers to checking whether intended actions were executed correctly and whether their results are correct; doing so could contribute to problem solving in both domains above by facilitating detection of errors (e.g., Ohlsson, 1996; Ohlsson & Rees, 1991). However, it is not well understood whether individual differences in planning and evaluating are stable across domains and tasks in math. The present study addressed this question.

### Allocation of Time

 Allocation of time refers to individuals’ determinations of how much time to spend on different tasks or items. Allocation of time could contribute to math problem solving in that individuals who allocate more time to more difficult items might perform better than individuals who spend equal amounts of time on items that vary in difficulty. This hypothesis was suggested in part by metareasoning research, which has found that individuals allocate time to items based on initial feelings of confidence (Ackerman & Thompson, 2017) and that individuals with better discrimination—that is, correspondence across items between feelings of confidence and accuracy—perform better on reasoning tasks (Jackson & Kleitman, 2014). The above hypothesis was also suggested by research on self-regulated study, which has found that learners study more the material they believe they know less well (Metcalfe, 2009; Morehead et al., 2017), and that learners who more accurately judge their own knowledge also learn better (Dunlosky & Rawson, 2012; Kornell & Metcalfe, 2006), including in math (Erickson & Heit, 2015).

 If, as these observations suggest, allocation of time contributes to math problem solving, is the contribution likely to be domain-general or domain-specific? It could be domain-general if it depends on traits such as executive function, which could facilitate assessment of item difficulty and enable deliberate rather than reactive allocation of time. However, assessment of item difficulty likely also depends on domain-specific knowledge, so the same individual might discriminate easy and hard items well in one domain and poorly in another. Consistent with this possibility, Dentakos et al. (2019) did not find that individual differences in discrimination (defined as in the previous paragraph) were correlated between different problem solving tasks. However, Dentakos et al. (2019) did not investigate whether differences in allocation of time are stable between tasks. We did so in the present study.

## The Present Study

 In the present study, we employed an individual differences paradigm to investigate relations of the above four skills to probabilistic reasoning and geometric proof justification. In principle, the four skills might vary among individuals in at least three ways: proficiency (how good one is at a skill), engagement (whether and how often one uses a skill), and quality of application (how well one uses a skill when one engages it). Whereas individual differences studies often focus on proficiency, the present study focused mainly on engagement, and secondarily on quality of application. Specifically, we sought evidence for engagement of our four candidate skills while solving math problems, then examined relations between problem solving accuracy and differences in engagement (and secondarily, quality of application). Strengths and limitations of this approach are considered in the Discussion.

 To address the above goal, we developed concurrent measures of engagement of the skills during performance of the PRS and GPJT. For logical reasoning, backward reference, and planning/evaluating, our measures were based on participants’ overt verbalizations, so we refer to these measures as “overt logical reasoning,” “overt backward reference,” and “overt planning/evaluating,” respectively. For allocation of time, our measures were based on correlations between response times and error rates, a proxy for item difficulty, so we refer to these measures as “RT-error correlations.” Each of these measures was taken separately for each task. We describe the measures in more detail under Method.

 We used the above measures to test three predictions for each of the four candidate skills. First, we reasoned that if a given skill contributes to performance on each task, then engagement of that skill while performing the PRS should be positively correlated with accuracy on the PRS, and engagement of that skill while performing the GPJT should be positively correlated with accuracy on the GPJT. Further, if differences in engagement of a skill are domain-general in nature, then engagement of the skill while performing the PRS should be correlated with engagement thereof while performing the GPJT. For a given skill, results supporting all of these predictions would be consistent with the hypothesis that the skill makes a domain-general contribution to math problem solving.

 Table 1 lists the predictions tested. Prediction 1 was that accuracies on the PRS and GPJT would be related, as in Braithwaite (2022). The other 12 predictions describe the three patterns of correlations explained above for each of the four candidate skills: logical reasoning (Predictions 2a-2c), backward reference (Predictions 3a-3c), planning/evaluating (Predictions 4a-4c), and allocation of time (Predictions 5a-5c).

==================== Insert Table 1 about here ====================

# Method

 This study was preregistered at <https://osf.io/uc3bw/?view_only=83860c17bcbf4e94866d03072c507e4c>. All deviations from the preregistration are noted below. Preregistered analyses are marked as such. Materials presented to participants, scripts followed by experimenters, and guidelines used for coding are provided in the Supplementary Materials.

## Participants

 Participants were undergraduate students recruited from the psychology department participant pool of Florida State University. We planned to recruit 107 participants, which would yield at least 85% power to detect moderate (*r* = .3) or larger correlations if no more than 10% of the sample were excluded, according to power analysis conducted in G\*Power 3.1 (Faul et al., 2007). Two extra participants signed up before recruitment was stopped, yielding an initial sample of 109. As preregistered, participants were excluded if they completed fewer than 75% of trials on either task (*n* = 6), if accuracy on either task did not statistically exceed chance (*n* = 0), or if they took more than two hours to complete the study (*n* = 0). Thus, the final sample included 103 participants. Seventy-four of these identified as female and 29 as male; 57 were first-year students; a variety of academic majors were reported, including psychology (26), biology (12), medical sciences (10), neuroscience (9), nursing (6), and exercise physiology (5).

Students in the US are expected to study statistics and probability in high school, including computation of probabilities of a union of events or of a joint event using the addition rule and multiplication rule (CCSSI, 2010). Similarly, US high school students are expected to study geometric proof, including proofs involving the three basic theorems for proving triangle congruence (SSS, SAS, ASA; *ibid.*). Consistent with the latter expectation, 102 participants in the analytical sample reported having taken a geometry course before and 96 reported that their geometry course included geometry proofs.

## Materials

 Participants completed the PRS and the GPJT in that order.

### Probabilistic Reasoning Scale

 The PRS was a modified version of the scale developed by Primi et al. (2017). The modified version, created by Braithwaite (2022), had the same items and answer choices as the original, but minor changes in wording. The PRS consisted of 16 multiple choice items, each with three answer choices. Primi et al. (2017) provided evidence for the scale’s validity. Cronbach’s alpha in this sample was .62.

 Items were displayed one at a time. To enable measurement of the time participants spent working on each question, excluding time spent reading the question, each question initially appeared without the answer choices, followed by a button labeled “Show Answers,” which participants were to press after reading the question. RT was measured as the time from pressing “Show Answers” until submission of a final answer.

### Geometric Proof Justification Task

 The GPJT (Braithwaite, 2022) consists of four proofs like the one shown in Figure 1, each consisting of 6 steps, for a total of 24 trials. The proofs involve similar diagrams, and all require using a triangle congruence theorem (e.g., Side-Angle-Side), but the proofs differ in what information is given and what claims are to be proven. Braithwaite (2022) provided evidence for the task’s validity. Cronbach’s alpha in this sample was .87.

 The GPJT was preceded by a review of basic geometry knowledge. This review covered terms and notations for line segments, angles, triangles, and congruence; definitions of congruent line segments, congruent angles, congruent triangles, bisection of line segments, bisection of angles, perpendicular lines; the principle that anything is congruent to itself; and three triangle congruence theorems (Side-Angle-Side, Angle-Side-Angle, and Side-Side-Side). After completing the geometry review, participants received a review sheet summarizing the above content, which they could refer to while performing the GPJT.

 Proofs were displayed one at a time. Initially, only the diagram, given information, and “to prove” statements were visible. After reading these, participants pressed a “Next” button, which revealed the first step of the proof, including the claim and the dropdown menu from which the justification for the claim was to be chosen. Upon choosing a justification, participants pressed “Next” again, which revealed the second step, and so on. Once a justification was chosen for the last step and “Next” was clicked, the proof disappeared and the next proof appeared. RT for each step was measured from the time “Next” was pressed to reveal the step to the time the “Next” was pressed again to reveal the next step or transition to the next proof.

## Procedure

 Participants completed the study in person under the supervision of an experimenter. All materials were presented on a desktop computer via Qualtrics. Participants completed all tasks in 40.7 minutes on average (min = 26.1, max = 70.4, *SD* = 9.7).

 First, participants were trained to think aloud following a script adapted from Fox et al. (2011). The script instructed participants to read each problem aloud and then say aloud all thoughts they had while working on the problem. They were instructed only to say their thoughts aloud and not to explain, because explaining has been shown to increase accuracy on cognitive tasks (Fox et al., 2011).[[2]](#footnote-3) Participants practiced thinking aloud with several simple problems that were unrelated to probability or geometry. Initially, some participants were hesitant to think aloud without prompting or long pauses, but after this practice, most began to do so and continued doing so throughout the following tasks.

 Next, participants completed the PRS, the geometry review, and the GPJT in that order. They were given paper and pencils for use as needed, but were instructed not to use calculators or other resources. They were instructed to think aloud while performing trials on both tasks, but not when reading instructions. During the geometry review, participants read instructional text silently but thought aloud while answering questions designed to test their recall of the content. After the PRS and GPJT, participants completed a short demographic survey.

## Measures of Concurrent Engagement of the Four Candidate Skills

 The protocol for each subject on each trial of each task was coded for each of three measures—overt logical reasoning, overt backward reference, and overt planning/evaluating—with 1 or 0 indicating presence or absence, respectively, of evidence in the protocol for engagement of the skill in question. All trials were coded independently by two coders, with all disagreements resolved through discussion. RT-error correlation was calculated across trials, separately for each task, without reference to the protocols. We describe each measure below.

### Overt Logical Reasoning

 Protocols were coded “1” for overt logical reasoning if the participant (1) used a logical connective word or phrase (e.g., “if”, “because,” “so,” “which means that”) and (2) this connective was used to express that one or more reasons implied a conclusion—for example, “I think there's more odd numbers **cause** it starts on an odd number and it ends on an odd number.” Uses of such words or phrases to connect calculations with their results (e.g., “three plus two, so five”), or as filler (e.g., “30 of the women smoke. So, um, it’s difficult”), were not counted for criterion (2). Such words or phrases that appeared while the participant was reading, rereading, or paraphrasing the problem statement or the geometry review sheet also were not counted. The coders agreed on whether these criteria were met for 79% of PRS trials and 85% of GPJT trials.

### Overt Backward Reference

 Protocols from the PRS were coded “1” for overt backward reference if the participant referred to the problem statement after reading it aloud for the first time. Protocols from the GPJT were coded “1” for backward reference if the participant referred to the problem statement after beginning to work on the first proof step, or to a previous step of the proof after beginning to work on the second proof step. The coders agreed on whether these criteria were met for 86% of PRS trials and 89% of GPJT trials.

### Overt Planning/Evaluating

 Protocols were considered to display planning if the participant described a plan or part of a plan to solve the problem or part of the problem before executing any steps of their solution[[3]](#footnote-4). Protocols were considered to display evaluating if the participant performed any operations or reasoning after stating an initial answer for the trial, reflecting an assumption that the primary reason for doing so would be a desire to check or verify the initial answer. Trials were coded “1” for overt planning/evaluating if they met the criteria for either planning or evaluating. The coders agreed on whether there was overt evidence of planning/evaluating on 86% of PRS trials and 91% of GPJT trials.

### RT-Error Correlation

 Separately for each task and each participant, we calculated the correlation between the amount of time the participant spent on each item and the entire sample’s error rate on each item. We refer to this measure as “RT-error correlation.” We reasoned that higher error rates on an item indicate higher item difficulty, so large positive RT-error correlations should be adaptive. As detailed below, the data were consistent with this assumption, in that for each task, RT-error correlations were positive on average, and individual differences in RT-error correlation were positively correlated with differences in accuracy.

# Results

 Means and SDs for all measures on both tasks are shown in Table 2.

==================== Insert Table 2 about here ====================

## Preregistered Analyses

 As predicted (Table 1, Prediction 1), individual differences in accuracies on the PRS and GPJT were positively correlated, *r*(101) = .51, *p* < .001. This result replicated Braithwaite (2022). Results of correlation analyses that tested the other predictions are shown in Figure 2. Except where indicated, significant correlations between a skill measure on a task and accuracy on the task remained significant when controlling for the other three skill measures on the task.

==================== Insert Figure 2 about here ====================

 As shown in Figure 2A, frequency of overt logical reasoning predicted accuracy on the GPJT (Prediction 2b) but not the PRS (Prediction 2a). Further, frequency of overt logical reasoning was correlated across tasks (Prediction 2c). Thus, individual differences in engagement of logical reasoning were somewhat stable across tasks, and such differences related to differences in geometric proof justification skill, but evidence was not found that such differences relate to differences in probabilistic reasoning skill.

 As shown in Figure 2B, frequency of overt backward reference correlated with accuracy on the GPJT (Prediction 3b) but not the PRS (Prediction 3a). The correlation between backward reference on two tasks did not reach significance (Prediction 3c). Thus, differences in backward reference were related to geometric proof justification skill, but were not found to be stable across tasks and or to be related to probabilistic reasoning skill.

 As shown in Figure 2C, frequency of overt planning/evaluating did not correlate with accuracy on either task (Predictions 4a and 4b). However, planning/evaluating was correlated across tasks (Prediction 4c). Thus, individual differences in frequency of engaging in these activities were somewhat stable across tasks, but no evidence was found that such differences contributed to competence on either task.

 Finally, as shown in Figure 2D, RT-error correlation on each task was correlated with accuracy on the same task (Predictions 5a and 5b)[[4]](#footnote-5). However, no evidence was found that RT-error correlations on the two tasks were related (Prediction 5c). Thus, the results were consistent with effective allocation of time contributing to competence on each task, but did not show that individual differences in allocation of time are stable across tasks.

## Exploratory Quantitative Analyses

 If our candidate higher-level skills accounted, in part, for the relation between accuracies on the PRS and GPJT, then the magnitude of that relation should be decreased by controlling for our concurrent measures of the candidate skills[[5]](#footnote-6). To test this possibility, for each candidate skill, we compared the correlation between PRS accuracy and GPJT accuracy to the analogous partial correlations obtained by regressing out both measures of the candidate skill, using Steiger’s (Steiger, 1980) test. For overt logical reasoning, the resulting partial correlation (*r* = .43) was smaller than the original (*r* = .51), *z* = 2.09, *p* = .036. The partial and original correlations did not differ when controlling for the other three concurrent skill measures, *p*s > .05.

 Overt logical reasoning, overt backward reference, and overt planning/evaluating were coded for individual trials, but the correlation analyses between these measures and accuracy (Figure 2) involved aggregating across trials. Thus, those analyses tested whether participants who engaged a skill more often were more accurate overall, but did not test whether engaging a skill on individual trials was associated with answering correctly on those trials. To address the latter question, we conducted a series of mixed logistic regressions with accuracy on individual trials (i.e., correct or not) as the dependent variable, one of the three measures above as a fixed effect, and participant as a random effect. The item presented on the trial was also included as a fixed effect. Six regressions were conducted, one per combination of the two tasks and the three measures. The regressions were conducted in *R* using *glmer* from *lme4* (Bates et al., 2014) with *p* values obtained using *lmerTest* (Kuznetsova et al., 2016).

 The logistic regressions with overt logical reasoning as a predictor found a positive effect for accuracy on the PRS, *B* = 0.39, *OR* = 1.48, *z* = 2.1, *p* = .040[[6]](#footnote-7), but not for accuracy on the GPJT, *B* = -0.19, *OR* = 0.83, *z* = -1.34, *p* = .18. Thus, in the PRS, participants were more likely to answer correctly on trials in which they displayed overt logical reasoning, but this was not the case for the GPJT. These findings contrast with Figure 1A, which shows—aggregating across trials—a relation between overt logical reasoning and accuracy for the GPJT but not the PRS.

 In contrast, results of the logistic regressions on overt backward reference and overt planning/evaluating were consistent with the earlier correlation analyses. Specifically, analogous to Figure 1B, overt backward reference did not predict accuracy in the PRS, *B* = -0.14, *OR* = 0.87, *z* = -0.67, *p* = .50, but did predict accuracy in the GPJT, *B* = 0.46, *OR* = 1.58, *z* = 3.26, *p* = .001. Analogous to Figure 1C, overt planning/evaluating did not predict accuracy in the PRS, *B* = 0.12, *OR* = 1.13, *z* = 0.61, *p* = .54, or the GPJT, *B* = -0.18, *OR* = 0.83, *z* = -1.30, *p* = .19.

## Exploratory Qualitative Analyses

 Quantitative analyses found relations between frequencies of overt logical reasoning on the two tasks, and between overt logical reasoning and accuracy on each task, albeit in different analyses for the PRS and GPJT. These findings are consistent with the possibility of a domain-general contribution of logical reasoning to math problem solving. To better understand the behaviors that produced the above relations, we conducted qualitative analyses of participants’ think-aloud protocols.

 First, for each task, we selected the six items that we calculated to have contributed most to the relations between overt logical reasoning and accuracy that were found in the quantitative analyses. We describe the selection process separately for the PRS and GPJT below. Next, for each selected item, we grouped participants’ protocols into four categories representing the four combinations of two variables: whether the protocol was coded as displaying overt logical reasoning, and whether the participant’s final answer was correct. Finally, we compared and contrasted the protocols within each of these groups.

 For protocols that were coded as displaying overt logical reasoning, we also analyzed their arguments using the framework of Toulmin (2007), as applied to mathematical reasoning by Jeannotte and Kieran (2017). According to this framework, an argument consists of (at least) three parts: evidence, warrant, and conclusion. In the present context, the evidence would be facts already known (or believed) about the problem at hand, the conclusion would be a claim asserted to be implied by the evidence, and the warrant would be a reason why the evidence implies the conclusion. In the PRS, the conclusion was often the final answer to the problem. In the GPJT, the conclusion was typically the claim to be justified in the current proof step, in which case the warrant was the answer to the problem (i.e., the justification selected by the participant). We identified evidence, warrant, and conclusion for each protocol that was coded as displaying overt logical reasoning. Then, we determined whether these jointly constituted a logically valid argument.

 We did not conduct qualitative protocol analyses relating to overt backward reference or overt planning/evaluating because our quantitative analyses did not find compelling evidence for domain-general contributions of these skills. Specifically, overt backward reference was not correlated across tasks or related to accuracy on the PRS, and overt planning/evaluating was not related to accuracy on either task.

### Overt Logical Reasoning in the Probabilistic Reasoning Scale

 For each item in the PRS, we calculated the difference in accuracy between participants who did or did not display overt logical reasoning on the item. We reasoned that the larger this difference, the larger the contribution of the item to the trial-level relation between overt logical reasoning and accuracy that was found in our mixed logistic regression. Thus, the six items with the largest difference scores were selected for qualitative analysis.

 The conclusions of our qualitative analysis are described below. The conclusions are based on analyses of all six items selected as described above. However, to illustrate the conclusions, we use example protocols from the two items with the largest difference scores, which are shown in Tables 3 and 4. Similar example tables for the other four selected items, and difference scores for all 16 items in the PRS, are provided in the Supplementary Materials. These items were similar to the remaining 10 items with respect to accuracy (87% vs. 86%) and frequency of overt logical reasoning (71% vs. 67%). The selected items involved calculating probabilities or frequencies (e.g., Table 3) or determining which of two outcomes was more likely (Table 4); the frequencies of these two types of items was similar among the selected items (4 and 2 respectively) and among the remaining items (6 and 4 respectively).

==================== Insert Tables 3 and 4 about here ====================

 **1. Participants used logical reasoning for a variety of purposes.** A common purpose was directly to answer the problem at hand, as illustrated in all four protocols in Table 4. Logical reasoning was also used to determine calculation(s) that needed to be performed in order to answer the problem, as illustrated in protocols 1, 2, and 4 of Table 3. Other uses of logical reasoning included estimating the answer or determining a plausible range for the answer (e.g., protocol 3 in Table 3), rejecting tempting but incorrect answers (e.g., protocol 2 in Supplementary Materials Table E5), and facilitating arithmetic calculations or algebraic manipulations (e.g., protocol 2 in Supplementary Materials Table E6).

 **2. The warrants of participants’ arguments were nearly always implicit.** Sometimes, the conclusion was also implicit. However, the implicit components of participants’ arguments were usually clear from context. For example, in protocol 2 in Table 3, the participant states that male smokers and female smokers represent 30% and 12% of the population respectively, which constitutes the evidence of their argument. The participant then states this conclusion: “so, we get 30 plus 12%.” The connection between evidence and conclusion depends on a basic principle of probability: if two events are mutually exclusive, the probability that one of them will occur is the sum of their probabilities. This principle constitutes the warrant of the argument; while not stated explicitly, it can be inferred as the most likely basis for the stated conclusion. Tables 3 and 4 display below each protocol the evidence, warrant, and conclusion that we inferred for it.

 **3. Participants’ arguments were virtually always logically valid.** In other words, the arguments’ conclusions logically followed from their premises[[7]](#footnote-8)—although the premises themselves, and therefore the conclusions, were not always true, as elaborated in point 4 below. This observation is illustrated by all protocols in Tables 3 and 4. The validity of participants’ arguments helps to explain the positive relation that was found in the mixed logistic regression between overt logical reasoning and accuracy on the PRS.

 **4. Corollary: When participants answered incorrectly despite using logical reasoning, the cause was not invalid logic.** Instead, the main causes of error in such cases were false evidence, false warrants, or errors occurring after logical reasoning. False evidence (e.g., protocol 3 in Table 3 and protocols 3 and 4 in Table 4) resulted from misreading or misunderstanding the problem statement, making incorrect assumptions, or committing calculation errors whose output served as evidence in an argument. False warrants included incorrect rules for determining marginal or joint probabilities (e.g., protocol 4 in Table 3) and well-documented probabilistic reasoning fallacies (e.g., gambler’s fallacy and hot hand fallacy; see protocol 4 of Supplementary Materials Table E5). Errors occurring after logical reasoning included misinterpreting the implications of a true conclusion regarding how to solve the problem (e.g., protocol 3 in Supplementary Materials Table E3), or correctly determining what calculation was needed via logical reasoning but performing the calculation incorrectly (e.g., protocol 4 in Supplementary Materials Table E4).

 **5. Protocols that did not receive the code for overt logical reasoning usually contained similar content to protocols that did receive this code.** In other words, these two types of protocols contained similar statements, calculations, and so on. The main difference between them related not to the content of individual statements, but rather to whether logical connective words and phrases were used to express logical relations between statements. This observation suggests that the positive correlation found between accuracy and overt logical reasoning on the PRS did not reflect qualitatively different thoughts being expressed by individuals who answered correctly and incorrectly, but instead reflected different levels of explicit attention to the logical relations among those thoughts.

### Overt Logical Reasoning in the Geometric Proof Justification Task

 As for the PRS, we identified items on the GPJT that contributed most to the relation between overt logical reasoning and accuracy. Because that relation was found at the level of participants rather than trials (Figure 2A), we ranked items based on the difference in accuracy between participants whose frequency of overt logical reasoning on the task was above or below median. We selected the six items with the largest difference scores for qualitative analysis.

 Also as for the PRS, our conclusions were based on qualitative analyses of all six selected items, but we illustrate the conclusions using protocols from the two items with the largest difference scores (Tables 5 and 6). These items were the fourth and fifth steps of the proof shown in Figure 1. Similar example tables for the other four selected items, and difference scores for all 24 items in the GPJT, are provided in the Supplementary Materials. The six selected items, compared to the remaining 18 items, elicited fewer correct responses (43% vs. 68%) but more displays of overt logical reasoning (45% vs. 30%). The selected items required justifying a claim of triangle congruence (e.g., Table 5) or drawing conclusions based on a previously-established triangle congruence (e.g., Table 6); the frequencies of these two types was 2 and 4, respectively, among the selected items and was 2 and 8, respectively, among the remaining items (the other 8 remaining items involved establishing the preconditions to use a triangle congruence theorem).

==================== Insert Tables 5 and 6 about here ====================

 **1. Participants used logical reasoning mainly to determine the correct answer.** More specifically, when participants displayed overt logical reasoning on a proof step in the GPJT, they usually constructed an argument whose conclusion was the claim to be justified in the current proof step, and whose warrant was the justification selected by the participant. This point contrasts with the more diverse uses of logical reasoning observed on the PRS (conclusion 1 in the previous subsection). It also contrasts with the frequent reliance on implicit warrants observed on the PRS (conclusion 2 in the previous subsection). The format of the GPJT essentially forced warrants to be explicit because they were selected as answers on the trials.

 **2. Both correct and incorrect answers were sometimes supported by logically valid reasoning,** as on the PRS. Correct answers supported by valid reasoning are illustrated by protocol 1 in each of Tables 5 and 6. In protocol 1 in Table 5 the participant argues that because two pairs of angles and one pair of sides are congruent (evidence), by the Angle-Side-Angle theorem (warrant), the triangles in question are congruent (conclusion). Similarly, in protocol 1 in Table 6 the participant argues that because the triangles have been shown to be congruent (evidence), by the definition of congruent triangles (warrant), their corresponding sides are congruent (conclusion).

 Incorrect answers supported by valid reasoning typically involved flawed evidence. These arguments were valid in the sense that the stated evidence and warrant jointly implied the conclusion, but the evidence was flawed in the sense of being either false or, more commonly, true but not proven to be true. For example, in protocol 3 in Table 5 the participant argues that because two pairs of sides and one pair of angles are congruent (evidence), by the Side-Angle-Side theorem (warrant), triangles BAM and CAM are congruent (conclusion). The flawed evidence is the claim that two pairs of sides are congruent; one of these pairs of sides (BM and CM) has not been proven congruent at this point in the proof. As another example, in protocol 3 in Table 6 the participant argues that because congruence of segments AB and AC is given in the problem statement (evidence), these segments are congruent (conclusion) by virtue of being given (warrant). The flawed evidence here is the false claim that congruence of AB and AC is given; in fact, the statement that AB and AC are congruent appears in the “to prove” part of the problem (Figure 1).

 **3. Logically invalid arguments were fairly common and accompanied both correct and incorrect answers.** This point contrasts with findings from the PRS, in which invalid arguments were vanishingly rare (conclusions 3 and 4 in the previous subsection). The frequency of invalid arguments on the GPJT may explain why overt logical reasoning did not predict correct responding on individual trials in this task.

 When correct answers were accompanied by invalid reasoning, participants chose the correct warrant for the proof step at hand, but failed explicitly to establish the conditions necessary for this warrant to be applicable. This type of argument is illustrated by protocol 2 in Tables 5 and 6. In protocol 2 in Table 5 the participant justifies the claim of triangle congruence (conclusion) by the Angle-Side-Angle theorem (warrant) and the fact that two pairs of angles are congruent (evidence). This argument is invalid because the evidence does not satisfy the conditions for applying the warrant, which also requires a pair of congruent sides. In protocol 2 in Table 6 the participant justifies the claim that segments AB and AC are congruent (conclusion) by the definition of congruent triangles (warrant) because AB and AC are sides of the same triangle (evidence). Again, the evidence cited does not license application of the warrant, making the argument invalid.

 Incorrect answers supported by invalid reasoning involved several sorts of logical error. First, these arguments sometimes involved choosing a warrant that could, in principle, imply the claimed conclusion (e.g., choosing a triangle congruence theorem to support a conclusion that two triangles are congruent), but without establishing that the conditions for applying this warrant were met. For example, in protocol 4 in Table 5 the participant makes almost the same argument as protocol 3 of the same table, except that only one pair each of congruent sides and angles are adduced as evidence, whereas the warrant (the Side-Angle-Side theorem) requires a second pair of congruent sides. A second sort of logical error that led to incorrect answers was selecting a warrant that could not, in principle, imply the claimed conclusion. For example, in protocol 4 of Table 6 the participant argues that AB and AC are congruent (conclusion) by virtue of the definition of bisecting an angle (warrant) because AB and AC are sides of the bisected angle BAC (evidence). This argument is invalid because, regardless of the evidence, the definition of bisecting an angle cannot justify a claim that two line segments are congruent[[8]](#footnote-9). A third sort of logical error was circular reasoning (e.g., protocol 4 in Supplementary Materials Table F4), in which the same claim serves as both evidence and conclusion[[9]](#footnote-10).

 **4. Protocols that were not coded as displaying overt logical reasoning most often involved simply stating an answer and nothing else.** This result, which contrasts with the verbose protocols recorded on the PRS even when participants did not display overt logical reasoning (conclusion 5 of the previous subsection), suggests that the skills on which participants relied to perform the GPJT may not always have been verbally mediated. Notably, protocols in which participants simply stated their answer and nothing else, and therefore were not coded as displaying overt logic reasoning, frequently accompanied both correct and incorrect answers. Such correct answers may reflect highly proficient individuals having automatized their thought process to a degree that verbal mediation was unnecessary, whereas the incorrect ones may reflect less proficient individuals lacking sufficient knowledge or understanding of geometric proof to generate coherent verbal arguments.

# Discussion

 The present study investigated the roles of four higher-level psychological skills—logical reasoning, backward reference, planning and evaluating, and allocation of time—in problem solving tasks in two mathematical domains, probability and geometry. Engagement of each skill during each task was measured, and relations of these measures to each other and to accuracy on each task were assessed. These analyses yielded evidence of domain-general contributions to problem solving for one of the candidate skills, logical reasoning, but not for the other three. Qualitative analyses of think-aloud protocols revealed similarities and differences in how participants used logical reasoning on the two tasks. Below, we discuss in detail implications of the findings regarding the role of logical reasoning in math problem solving, then more briefly implications regarding roles of the other three candidate skills.

## The Role of Logical Reasoning in Math Problem Solving

 Numerous previous studies have found predictive relations between logical reasoning and math problem solving, even when controlling for various possible confounds such as fluid intelligence, number line estimation, arithmetic skill, and algebra skill (Battista, 1990; Braithwaite, 2022; Kleemans et al., 2018; Morsanyi et al., 2018; Nunes et al., 2007; Wong, 2018). These findings suggest that logical reasoning may contribute to math problem solving. However, the above studies assessed the two constructs using separate tasks, and therefore did not provide direct evidence of logical reasoning being used during math problem solving. The concurrent measures of logical reasoning employed in the present study provided such evidence. Further, these measures—like the separate logical reasoning measures used in prior studies—predicted math problem solving accuracy.

 Our reliance on concurrent measures of logical reasoning enables us to discriminate between two hypotheses about how logical reasoning could contribute to math problem solving: an offline hypothesis and an online hypothesis. The offline hypothesis is that logical reasoning facilitates acquisition of problem solving skills, so that individuals who are strong in logical reasoning tend to be better problem solvers. The online hypothesis is that logical reasoning during problem solving leads to more accurate performance. The previous findings cited above are equally well explained by either hypothesis, but only the online hypothesis implies an association between accuracy on problem solving tasks and concurrent logical reasoning on the same tasks. Thus, the present findings are better explained by the online hypothesis. Of course, the hypotheses are not exclusive, so evidence in favor of one does not undermine the other.

 Our quantitative measures of logical reasoning assessed whether and how often logical reasoning was overtly displayed, without regard to whether the reasoning was correct. Thus, it may seem surprising that these measures were related to problem solving accuracy. However, this finding is less surprising when we consider how making one’s reasoning explicit could constrain the problem solving process. When reasoning is left implicit, one may make claims freely without considering whether or how well they can be supported by reasons. Making reasoning explicit requires providing reasons for one’s conclusions. This constraint could cause individuals to avoid claims that they cannot support well in favor of ones that they can, which could improve accuracy even among poor reasoners, provided that they have at least some ability—however imperfect—to distinguish between strong reasons and weak ones.

### The Role of Logical Reasoning on the PRS

 On the PRS, participants were more likely to answer correctly on trials during which they displayed logical reasoning. Overt logical reasoning did not guarantee a correct answer, but when participants erred despite displaying logical reasoning, their errors were virtually never *logical* errors. Far more often, the fault was in the premises of their arguments: they appealed to evidence that was not accurate or relied on warrants that were not mathematically sound. These errors appear to reflect weaknesses in mathematical knowledge rather than logic.

 Why were logical errors so rare on the PRS, whereas such errors have appeared pervasive in many studies of logical reasoning (Khemlani & Johnson-Laird, 2012; Oberauer, 2006; Wason, 1968)? A possible explanation is that participants usually employed logical forms with which people rarely err. This point is well illustrated by protocol 2 in Table 4, whose argument could be paraphrased as “If there are more odd numbers, then an odd number is more likely; there are more odd numbers; therefore an odd number is more likely.” The logical form of this argument, called *modus ponens*, elicits near-ceiling accuracy from college-educated adults (e.g., Morsanyi et al., 2018; Oberauer, 2006), whereas inference forms that elicit higher error rates (e.g., *modus tollens*: “If A then B, not B, therefore not A”) were not observed on the PRS. Thus, participants may have avoided logical errors by using primarily or only the easiest logical forms.

 Two previous studies found relations between accuracy on logical reasoning tasks and accuracy on the PRS (Braithwaite, 2022; Morsanyi et al., 2018). These findings could have been explained by assuming that logical reasoning skill affects both logical and probabilistic reasoning by affecting the frequency of logical errors. However, such an explanation is challenged by the extreme rarity of logical errors on the PRS in the present study. We therefore propose that logical reasoning skill affects logical and probabilistic reasoning accuracy via different mechanisms. On logical reasoning tasks, logical reasoning skill facilitates judgments of validity, thus directly improving accuracy. In probabilistic reasoning, logical reasoning skill increases the likelihood of engaging in logical reasoning, which in turn increases accuracy by constraining responses as described in the previous section. If this proposal is correct, then merely inducing individuals to use logic during probabilistic reasoning could improve accuracy without the need to improve the quality of their logical reasoning—a prediction worth testing in the future.

### The Role of Logical Reasoning on the GPJT

 On the GPJT, individuals who more frequently displayed overt logical reasoning also displayed higher accuracy. This relation could reflect differences in logical reasoning skill—that is, individuals who were more skilled in logical reasoning may have been more likely to engage in such reasoning while performing the GPJT, leading to higher accuracy. Alternatively, or in addition, the relation could reflect different attitudes towards the task. That is, some individuals might be more inclined than others to expect, and seek, logical relations between each proof step and the given information and/or previous proof steps; this expectation may have led to more frequent overt displays of logical reasoning and also to higher accuracy.

 A striking difference between the GPJT and PRS was the low frequency of overt logical reasoning on the former, compared to the latter (34% vs. 68%). Further, trials that were not coded as displaying overt logical reasoning often involved simply stating the answer and nothing else. These results suggest that answer selection on the GPJT often relied on processes that were not verbally mediated (such as spatial reasoning) and so unlikely to involve logic. Also consistent with this interpretation is the finding that overt use of logical reasoning was not associated with greater likelihood of a correct answer at the level of individual trials.

 Another striking difference between the tasks was the high frequency of logically invalid arguments on the GPJT, compared to the rarity of such arguments on the PRS. This result is surprising considering that geometric proof has been used as a primary context for teaching logical reasoning in math (Sinclair, 2008). One possible reason is that the GPJT asks students to prove things that are already known to be true. Some students may perceive such activities as meaningless (Hoyles, 1997), which could decrease the quality of their logical reasoning. Another possible reason relates to the two-column proof format used in the GPJT. This format is common in geometry textbooks and classrooms (Nirode & Boyd, 2021), but has been criticized for excessive emphasis on form at the expense of meaning (Sowder & Harel, 1998). Thus, the format might increase some students’ perception that the task is about something other than constructing logically coherent arguments—for example, guessing what goes in each blank. This point does not necessarily detract from the task’s utility as a measure of individual differences, but may explain why it elicited invalid reasoning more often than the PRS.

### Relations Between Logical Reasoning on the PRS and GPJT

 Logic can be described and studied in a formal manner—that is, without reference to content. However, human logical reasoning depends on content as well as form (Khemlani et al., 2018; Oaksford & Chater, 2020). For example, people judge arguments to be valid more often when the arguments’ conclusions accord with prior beliefs (Evans et al., 1983).

 This content-dependence of logical reasoning implies that individual differences in logical reasoning need not be stable across domains. Someone might reason logically in a domain about which they were knowledgeable, but not in another domain about which they were ignorant. However, in the present study, frequency of overt logical reasoning on the PRS was correlated with the analogous measure on the GPJT. Thus, individual differences in logical reasoning during math problem solving were at least somewhat stable across domains. Moreover, controlling for these differences significantly reduced the correlation between accuracies on the two tasks. These findings are consistent with our hypothesis that logical reasoning not only contributes to math problem solving but also does so in a domain-general manner.

## Other Candidates for Higher-Level Domain-General Skills in Math Problem Solving

### Backward Reference

 Findings from Braithwaite (2022) suggested that individual differences in the tendency to refer back to the problem statement and previous steps contributed to differences in performance on the GPJT. Consistent with this possibility, in the present study, accuracy on the GPJT was strongly (*r* = .49) correlated with frequency of backward reference on that task. However, an analogous association was not found on the PRS, nor was frequency of backward reference correlated across tasks. We speculate that, although understanding and elaborating problem statements is likely important for problem solving in any domain of math (e.g., Cummins et al., 1988; Hegarty et al., 1995; Nathan et al., 1992; Sebrechts et al., 1996), the phenomenon of neglecting the problem statement as a possible source of justifications for claims may be specific to geometric proof or even to the GPJT. In the GPJT, this phenomenon might be a consequence of some individuals approaching the task without an expectation of logical coherence as suggested above, or alternatively, of an incorrect belief that a claim being given does not constitute a legitimate justification for the claim.

### Planning and Evaluating

 Frequencies of two metacognitive activities, planning and evaluating, were correlated across tasks, indicating stable individual differences in tendencies to engage in these activities. This result dovetails with some previous studies that have found evidence for domain-generality of metacognitive skills, especially for older children (Bellon et al., 2020; Veenman & Spaans, 2005) and adults (Veenman et al., 1997).

 However, differences in frequency of planning and evaluating did not predict accuracy on either task. Benefits of planning may have been minimal on the PRS because that task involved many items that could be solved by rote execution of procedures, and on the GPJT because participants were not required to construct proofs but only to justify steps in them. Evaluating, on the other hand, may have conferred minimal benefits if evaluation was not habitual but instead contingent on cues that are correlated with the presence of errors, such as feelings of uncertainty. If so, then participants who frequently evaluated would not be expected to display higher accuracy than those who rarely evaluated due to high confidence in their solutions (see Braithwaite & Sprague, 2021 for evidence of this phenomenon). Both planning and evaluating might confer greater benefits in more complex problem solving tasks where individuals are unlikely to possess ready-made solution methods (e.g., Cai et al., 2016).

### Allocation of Time

 Findings from research on metareasoning and self-regulated learning (Ackerman & Thompson, 2017; Erickson & Heit, 2015) suggested that the ability to allocate time preferentially to more difficult items could contribute to competence in math problem solving. Consistent with this possibility, in both tasks, participants on average spent more time on items that elicited more errors, and the degree to which they did so—that is, the RT-error correlation—was positively related to performance on each task. However, RT-error correlations were not correlated across tasks. This null finding dovetails with a recent study of adults’ problem solving (Dentakos et al., 2019) in which participants solved problems involving probability calculation, financial calculation, and emotion recognition, and reported confidence in their answers after each trial. Within tasks, confidence ratings across trials were positively correlated with accuracy. However, these accuracy-confidence correlations—like RT-error correlations in the present study—were not correlated across tasks. Both that study and the present one suggest that individual differences in discrimination among items within a task are domain-specific, perhaps because assessment of item difficulty depends on specific knowledge of the domain.

## Limitations

 A limitation of this study is that our analyses did not control for individual differences in basic capacities, such as working memory or spatial ability, or relevant content knowledge, such as knowledge of geometry and probability. This limitation is partially addressed by our reliance on concurrent measures, which allowed us to observe—rather than infer—that the higher-level skills were engaged during the problem solving tasks, and by the fact that logical reasoning skill was previously found to correlate with accuracies on both the PRS and GPJT even when controlling for various other skills (Braithwaite, 2022; Morsanyi et al., 2018). Nevertheless, it remains possible that the relations that were found in the present study between various measures, such as frequency of overt logical reasoning and accuracies on the two problem solving tasks, were partially or entirely driven by differences in basic cognitive capacities and/or relevant content knowledge. It would be desirable in the future to control for these variables and thereby isolate variation that is both specifically higher-level and specifically domain-general.

 Another limitation is that the PRS and GPJT differed not only by involving different mathematical domains, but in many other ways as well. For example, the PRS required deriving initially unknown information from given information, whereas the GPJT required selecting justifications for statements known to be true. The numerous differences between the tasks renders it impossible to determine which ones account for the different findings obtained with the tasks—for example, the greater frequency of overt logical reasoning, but lower frequency of logically invalid arguments, on the PRS, and the fact that overt logical reasoning was associated with accuracy only on individual trials on the PRS, but only across trials on the GPJT. Future research should investigate how such task-level factors impact contributions of logical reasoning, and other potentially domain-general skills, to performance.

## Conclusion

 Individual differences in overt logical reasoning during math problem solving were correlated across mathematical tasks involving probability and geometry, and overt logical reasoning was related to accurate performance on both tasks. These findings suggest that differences in overt logical reasoning contribute to differences math problem solving across domains. Evidence for an analogous conclusion was not found for three other higher-level skills—backward reference, planning and evaluating, and allocation of time. Results also pointed to task-related differences in how logical reasoning is used and how doing so contributes to accurate performance. In particular, we found that overt logical reasoning was more common on the PRS, whereas logically invalid reasoning was more common on the GPJT. We hope the findings will stimulate further investigation of factors that affect engagement of logical reasoning in math, mechanisms underlying relations between logical reasoning and accuracy, and how these mechanisms differ among different mathematical domains and tasks.

# References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Ahmadpour, F., Reid, D., & Reza Fadaee, M. (2019). Students’ ways of understanding a proof. *Mathematical Thinking and Learning*, *21*(2), 85–104. https://doi.org/10.1080/10986065.2019.1570833

Attridge, N., Doritou, M., & Inglis, M. (2015). The development of reasoning skills during compulsory 16 to 18 mathematics education. *Research in Mathematics Education*, *17*(1), 20–37. https://doi.org/10.1080/14794802.2014.999014

Attridge, N., & Inglis, M. (2013). Advanced mathematical study and the development of conditional reasoning skills. *PLoS ONE*, *8*(7), e69399. https://doi.org/10.1371/journal.pone.0069399

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4*. *67*(1). https://doi.org/10.18637/jss.v067.i01

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, *21*(1), 47–60. https://doi.org/10.2307/749456

Bellon, E., Fias, W., & de Smedt, B. (2020). Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific? *PLOS ONE*, *15*(3), e0229932. https://doi.org/10.1371/JOURNAL.PONE.0229932

Braithwaite, D. W. (2022). Relations between geometric proof justification and probabilistic reasoning. *Learning and Individual Differences*, *98*, 102201. https://doi.org/10.1016/j.lindif.2022.102201

Braithwaite, D. W., & Rafferty, A. N. (n.d.). *Knowledge of Examples Affects Conditional Reasoning with Mathematical Content*.

Braithwaite, D. W., & Sprague, L. (2021). Conceptual knowledge, procedural knowledge, and metacognition in routine and nonroutine problem solving. *Cognitive Science*, *45*, e13048. https://doi.org/10.1111/cogs.13048

Cai, D., Georgiou, G. K., Wen, M., & Das, J. P. (2016). The role of planning in different mathematical skills. *Journal of Cognitive Psychology*, *28*(2), 234–241. https://doi.org/10.1080/20445911.2015.1103742

Cirillo, M., & Hummer, J. (2021). Competencies and behaviors observed when students solve geometry proof problems: an interview study with smartpen technology. *ZDM – Mathematics Education*, *53*(4), 861–875. https://doi.org/10.1007/s11858-021-01221-w

Clements, D. H., Wilson, D. C., & Sarama, J. (2004). Young Children’s Composition of Geometric Figures: A Learning Trajectory. *Mathematical Thinking and Learning*, *6*(2), 163–184. https://doi.org/10.1207/s15327833mtl0602\_5

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: the Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25–47. https://psycnet.apa.org/record/2012-03055-003

Cornoldi, C., Carretti, B., Drusi, S., & Tencati, C. (2015). Improving problem solving in primary school students: The effect of a training programme focusing on metacognition and working memory. *British Journal of Educational Psychology*, *85*(3), 424–439. https://doi.org/10.1111/bjep.12083

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4

Dentakos, S., Saoud, W., Ackerman, R., & Toplak, M. E. (2019). Does domain matter? Monitoring accuracy across domains. *Metacognition and Learning*, *14*(3), 413–436. https://doi.org/10.1007/S11409-019-09198-4/TABLES/6

Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem solving? *Journal of Educational Psychology*, *95*(1), 188–200. https://doi.org/10.1037/0022-0663.95.1.188

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students’ learning and retention. *Learning and Instruction*, *22*(4), 271–280. https://doi.org/10.1016/J.LEARNINSTRUC.2011.08.003

Erickson, S., & Heit, E. (2015). Metacognition and confidence: comparing math to other academic subjects. *Frontiers in Psychology*, *6*, 742. https://doi.org/10.3389/fpsyg.2015.00742

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*(3), 295–306. https://doi.org/10.3758/BF03196976/METRICS

Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition*, *35*(7), 1772–1784.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Garofalo, J., & Lester, F. K. Jr. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education*, *16*(3), 163–176. https://doi.org/10.5951/jresematheduc.16.3.0163

Glenberg, A. M., Willford, J., Gibson, B., Goldberg, A., & Zhu, X. (2012). Improving reading to improve math. *Scientific Studies of Reading*, *16*(4), 316–340. https://doi.org/10.1080/10888438.2011.564245

Hacker, D. J., Kiuhara, S. A., & Levin, J. R. (2019). A metacognitive intervention for teaching fractions to students with or at-risk for learning disabilities in mathematics. *ZDM*, *51*(4), 601–612. https://doi.org/10.1007/s11858-019-01040-0

Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. K. , Jr. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 805–842). National Council of Teachers of Mathematics.

Hawes, Z. C. K., Gilligan-Lee, K. A., & Mix, K. S. (2022). Effects of Spatial Training on Mathematics Performance: A Meta-Analysis. *Developmental Psychology*, *58*(1), 112–137. https://doi.org/10.1037/DEV0001281

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*(1), 18–32. https://doi.org/10.1037/0022-0663.87.1.18

Herbst, P. G., Fujita, T., Halverscheid, S., & Weiss, M. (2017). *The learning and teaching of geometry in secondary schools: A modeling perspective*. Routledge.

Hoyles, C. (1997). The Curricular Shaping of Students’ Approaches to Proof. *For the Learning of Mathematics*, *17*(1), 7–16.

Inglis, M., & Simpson, A. (2009). Conditional inference and advanced mathematical study: further evidence. *Educ Stud Math*, *72*, 185–198. https://doi.org/10.1007/s10649-009-9187-z

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*(1), 25–49. https://doi.org/10.1007/S11409-013-9110-Y/FIGURES/3

Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics*, *96*(1), 1–16. https://doi.org/10.1007/s10649-017-9761-8

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, *50*(1), 109–135. https://doi.org/10.1146/annurev.psych.50.1.109

Khemlani, S., & Johnson-Laird, P. N. (2022). Reasoning about properties: A computational theory. *Psychological Review*, *129*(2), 289–312. https://doi.org/10.1037/rev0000240

Khemlani, S. S., Byrne, R. M. J., & Johnson‐Laird, P. N. (2018). Facts and possibilities: A model‐based theory of sentential reasoning. *Cognitive Science*, *42*(6), 1887–1924. https://doi.org/10.1111/cogs.12634

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457. https://doi.org/10.1037/a0026841

Kleemans, T., Segers, E., & Verhoeven, L. (2018). Role of linguistic skills in fifth-grade mathematics. *Journal of Experimental Child Psychology*, *167*, 404–413. https://doi.org/10.1016/J.JECP.2017.11.012

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, *14*(4), 511–550.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://psycnet.apa.org/journals/xlm/32/3/609/

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in Linear Mixed Effects Models* (2.0-33).

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Mevarech, Z. R., & Kramarski, B. (2003). The effects of metacognitive training versus worked-out examples on students’ mathematical reasoning. *British Journal of Educational Psychology*, *73*(4), 449–471. https://doi.org/10.1348/000709903322591181

Morehead, K., Dunlosky, J., & Foster, N. L. (2017). Do people use category-learning judgments to regulate their learning of natural categories? *Memory & Cognition*, *45*(8), 1253–1269. https://doi.org/10.3758/s13421-017-0729-9

Morsanyi, K., McCormack, T., & O’Mahony, E. (2018). The link between deductive reasoning and mathematics. *Thinking & Reasoning*, *24*(2), 234–257. https://doi.org/10.1080/13546783.2017.1384760

Muncer, G., Higham, P. A., Gosling, C. J., Cortese, S., Wood-Downie, H., & Hadwin, J. A. (2022). A Meta-Analysis Investigating the Association Between Metacognition and Math Performance in Adolescence. *Educational Psychology Review*, *34*(1), 301–334. https://doi.org/10.1007/S10648-021-09620-X/TABLES/3

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, *9*(4), 329–389. https://doi.org/10.1207/s1532690xci0904\_2

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Authors.

Newell, A., & Simon, H. A. (1972). *Human problem solving: Vol. 140.9*. Prentice-Hall. http://www.sci.brooklyn.cuny.edu/~kopec/cis718/fall\_2005/2/Rafique\_2\_humanthinking.doc

Nirode, W., & Boyd, B. (2021). High school geometry textbooks’ proving opportunities of Common Core theorems. *School Science and Mathematics*, *121*(6), 345–356. https://doi.org/10.1111/ssm.12487

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 398–415. http://www.ncbi.nlm.nih.gov/pubmed/1829473

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A., & Carraher, J. (2007). The contribution of logical reasoning to the learning of mathematics in primary school. *British Journal of Developmental Psychology*, *25*(1), 147–166. https://doi.org/10.1348/026151006X153127

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*(1), 305–330. https://doi.org/10.1146/annurev-psych-010419-051132

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, *53*(3), 238–283. https://doi.org/10.1016/J.COGPSYCH.2006.04.001

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review*, *103*(2), 241–262. https://doi.org/10.1037/0033-295X.103.2.241

Ohlsson, S., & Rees, E. (1991). The function of conceptual understanding in the learning of arithmetic procedures. *Cognition and Instruction*, *8*(2), 103–179. https://doi.org/10.1207/s1532690xci0802\_1

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473. https://doi.org/10.1037/edu0000079

Primi, C., Morsanyi, K., Donati, M. A., Galli, S., & Chiesi, F. (2017). Measuring probabilistic reasoning: The construction of a new scale applying Item Response Theory. *Journal of Behavioral Decision Making*, *30*(4), 933–950. https://doi.org/10.1002/bdm.2011

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). MacMillan.

Schoenfeld, A. H. (2009). The soul of mathematics. In D. Stylianou, M. Blanton, & E. Knuth (Eds.), *Teaching and learning proof across the grades: a K-16 perspective* (pp. xii–xvi). Taylor & Francis Group.

Sebrechts, M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction*, *14*(3), 285–343. https://doi.org/10.1207/s1532690xci1403\_2

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Sinclair, N. (2008). *The history of the geometry curriculum in the United States*. Information Age Publishing.

Sinclair, N., & Moss, J. (2012). The more it changes, the more it becomes the same: The development of the routine of shape identification in dynamic geometry environment. *International Journal of Educational Research*, *51–52*, 28–44. https://doi.org/10.1016/j.ijer.2011.12.009

Sobkow, A., Olszewska, A., & Traczyk, J. (2020). Multiple numeric competencies predict decision outcomes beyond fluid intelligence and cognitive reflection. *Intelligence*, *80*, 101452. https://doi.org/10.1016/J.INTELL.2020.101452

Sommerhoff, D., & Ufer, S. (2019). Acceptance criteria for validating mathematical proofs used by school students, university students, and mathematicians in the context of teaching. *ZDM*, *51*(5), 717–730. https://doi.org/10.1007/s11858-019-01039-7

Sowder, L., & Harel, G. (1998). Types of students’ justifications. *The Mathematics Teacher*, *91*(8), 670–675. https://doi.org/10.5951/MT.91.8.0670

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stillman, G. A., & Galbraith, P. L. (1998). Applying mathematics with real world connections: metacognitive characteristics of secondary students. *Educational Studies in Mathematics*, *36*(2), 157–194. https://doi.org/10.1023/A:1003246329257

Stylianides, G. J., Stylianides, A. J., & Weber, K. (2017). Research on the teaching and learning of proof: Taking stock and moving forward. In J. Cai (Ed.), *Compendium for Research in Mathematics Education* (pp. 237–266). National Council of Teachers of Mathematics.

Szilágyi, J., Clements, D. H., & Sarama, J. (2013). Young Children’s Understandings of Length Measurement: Evaluating a Learning Trajectory. *Journal for Research in Mathematics Education*, *44*(3), 581–620. https://doi.org/10.5951/jresematheduc.44.3.0581

Thompson, V. A., & Markovits, H. (2021). Reasoning strategy vs cognitive capacity as predictors of individual differences in reasoning performance. *Cognition*, *217*, 104866. https://doi.org/10.1016/J.COGNITION.2021.104866

Toulmin, S. E. (2007). *The uses of argument*. Cambridge University Press.

Tversky, A., & Kahneman, D. (1993). Probabilistic reasoning. In *Readings in philosophy and cognitive science* (pp. 43–68). https://books.google.com/books?hl=en&lr=&id=mv2vGDT-6KIC&oi=fnd&pg=PA43&dq=probabilistic+reasoning+intuitive+OR+impulsive+error&ots=4HdfSnz36D&sig=G\_3dJIRo4N3yXbsNxQG36RPKqx0

Veenman, M. V. J., Elshout, J. J., & Meijer, J. (1997). The generality vs domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction*, *7*(2), 187–209. https://doi.org/10.1016/S0959-4752(96)00025-4

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*(2), 159–176. https://doi.org/10.1016/J.LINDIF.2004.12.001

Veenman, M. V. J., & van Cleef, D. (2019). Measuring metacognitive skills for mathematics: students’ self-reports versus on-line assessment methods. *ZDM*, *51*(4), 691–701. https://doi.org/10.1007/s11858-018-1006-5

Wason, P. C. (1968). Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wong, T. T. Y. (2018). Is conditional reasoning related to mathematical problem solving? *Developmental Science*, *21*(5), 1–12. https://doi.org/10.1111/desc.12644

Wu, H. H. (1996). The role of Euclidean geometry in high school. *Journal of Mathematical Behavior*, *15*(3), 221–237. https://doi.org/10.1016/S0732-3123(96)90002-4

Yang, K.-L. (2012). Structures of cognitive and metacognitive reading strategy use for reading comprehension of geometry proof. *Educational Studies in Mathematics*, *80*(3), 307–326. https://doi.org/10.1007/s10649-011-9350-1

Yang, K.-L., & Lin, F.-L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, *67*(1), 59–76. https://doi.org/10.1007/s10649-007-9080-6

Zhao, N., Teng, X., Li, W., Li, Y., Wang, S., Wen, H., & Yi, M. (2019). A path model for metacognition and its relation to problem-solving strategies and achievement for different tasks. *ZDM*, *51*(4), 641–653. https://doi.org/10.1007/s11858-019-01067-3

# Figures and Tables



Figure 1. One proof in the Geometric Proof Justification Task, with correct answers displayed.



Figure 2. Results of correlation analyses testing Predictions 2a-5c in Table 1. Solid and dashed lines denote significant and non-significant correlations.

Table 1. Predictions Tested in the Present Study.

|  |  |
| --- | --- |
| Number | Prediction |
| 1 | Accuracies on the PRS and GPJT are positively correlated |
| 2a | Overt logical reasoning on the PRS is positively correlated with PRS accuracy |
| 2b | Overt logical reasoning on the GPJT is positively correlated with GPJT accuracy |
| 2c | Overt logical reasoning on the PRS and GPJT are positively correlated |
| 3a | Overt backward reference on the PRS is positively correlated with PRS accuracy |
| 3b | Overt backward reference on the GPJT is positively correlated with GPJT accuracy |
| 3c | Overt backward reference on the PRS and GPJT are positively correlated |
| 4a | Overt planning/evaluating on the PRS is positively correlated with PRS accuracy |
| 4b | Overt planning/evaluating on the GPJT is positively correlated with GPJT accuracy |
| 4c | Overt planning/evaluating on the PRS and GPJT are positively correlated |
| 5a | RT-error correlation on the PRS is positively correlated with PRS accuracy |
| 5b | RT-error correlation on the GPJT is positively correlated with GPJT accuracy |
| 5c | RT-error correlations on the PRS and GPJT are positively correlated |

Table 2. Means and SDs for All Measures on Both Tasks.

|  |  |  |
| --- | --- | --- |
|  | PRS | GPJT |
|  | Mean | SD | Mean | SD |
| (1) Accuracy | .86 | .13 | .62 | .21 |
| (2) Overt Logical Reasoning | .68 | .22 | .34 | .21 |
| (3) Overt Backward Reference | .23 | .18 | .42 | .19 |
| (4) Overt Planning/Evaluating | .30 | .18 | .21 | .13 |
| (5) RT-Error Correlation | .18 | .24 | .26 | .21 |

Table 3. Protocols for “60% of the population in a city are men and 40% are women. 50% of the men and 30% of the women smoke. We select a person from the city at random. What is the probability that this person is a smoker? (a) 42% (\*), (b) 50%, (c) 85%.”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Answer Correct? | Yes | Yes | No | No |
| Protocol | “Well, I think what you would do is, you would first say, alright, we know half the men, **so**, we know 30%, half the men, which makes up 60%. **So**, that's 30% of the population.” | “Men who smoke is going to be 60% times 50%, which is gonna be 30%. And you're gonna add that to the population of women who smoke, which is 40% times 30%, which is 12%. **So**, we get 30 plus 12% is 42%.” | “I would probably just choose 85 **since** it's the closest, more than half of each of them are being chosen, or are smokers. **So**, more than half of them would be smokers of the whole population.” | “50 plus 30% is 80. … **Because** it's person, not men or women, you would combine the two. **So**, 60 again, that makes up 100% of the population. **So**, of all the population, 80% of it smokes, 80% of the people in this population smokes.” |
| Evidence | 60% of the population are men and 50% of the men smoke. | 30% of the population are men who smoke and 12% of the population of women who smoke. | More than half of men are smokers and more than half of women are smokers. | 50% of men smoke and 30% of women smoke. |
| Warrant | P(A & B) = P(A) \* P(B|A) | If A and B are exclusive, P(A or B) = P(A) + P(B) | If P(C|A) > *x* and P(C|B) > *x*, then P(C|A or B) > *x* | If A and B are exclusive and P(C|A) = *x* and P(C|B) = *y*, then P(C|A or B) = *x* + *y* |
| Conclusion | The percentage of the population that are men who smoke is 60% \* 50%. | The percentage of the population that smoke is 30% + 12%. | More than half of the population are smokers. | The percentage of the population that smoke is 50% + 30%. |
| Evidence true? | Yes | Yes | No | Yes |
| Warrant true? | Yes | Yes | Yes | No |
| Valid? | Yes | Yes | Yes | Yes |

Note. Logical connective words in the protocols are **bolded**.

Table 4. Protocols for “A bingo game is played with 25 numbers (from 1 to 25). At the first draw, which of the following results is the most likely? (a) It is more likely to be an even number, (b) It is more likely to be an odd number (\*), (c) It is just as likely to be an even or an odd number.”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Answer Correct? | Yes | Yes | No | No |
| Protocol | “I think there's more odd numbers **cause** it starts on an odd number and it ends on an odd number.” | “I guess **if** there's whichever one is more of, even if it's by one number would be slightly more likely. … There's 1, 2, 3-- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 odd numbers. And then there are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 even numbers. **So**, I guess it's slightly more likely to be odd.” | “**Because** a bingo game is played with 25 numbers from 1 to 25, there's we pretty sure there's an equal amount of even and odd numbers in 1 to 25 after the first draw, which of the following results, it sounds like I'm gonna say it's likely to be an even or odd number.” | “… 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. There's 12 odd numbers. That'd make 13 even numbers **so** it'd most likely be an even number, I'll say.” |
| Evidence | The sequence 1 to 25 starts on an odd number and ends on an odd number. | The sequence 1 to 25 contains 13 odd numbers and 12 even numbers. | The sequence 1 to 25 contains equally many even and odd numbers. | The sequence 1 to 25 contains 12 odd numbers and 13 even numbers. |
| Warrant | If a sequence of successive integers starts and ends with the same parity, then it contains more numbers with that parity than the opposite. | If A and B are sets of equiprobable outcomes and |A| > |B|, then P(A) > P(B). | If A and B are sets of equiprobable outcomes and |A| = |B|, then P(A) = P(B). | If A and B are sets of equiprobable outcomes and |A| > |B|, then P(A) > P(B). |
| Conclusion | The sequence 1 to 25 contains more odd than even numbers. | It is more likely to be an odd number than an even number. | It is equally likely to be an even or odd number. | It is more likely to be an even number than an odd number. |
| Evidence true? | Yes | Yes | No | No |
| Warrant true? | Yes | Yes | Yes | Yes |
| Valid? | Yes | Yes | Yes | Yes |

Note. Logical connective words in the protocols are **bolded**.

Table 5. Protocols for the fourth step of the proof in Figure 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Answer Correct? | Yes | Yes | No | No |
| Protocol | “Triangle BAM and triangle CAM would be congruent **because, so,** BAM that angle M for both are proven and then well we proved that AM, AM, and then angle BMA and angle CMA. **So**, we have the angle side and then another angle. **So**, angle side angle.” | “Triangle BAM is congruent to CAM **because of** angle side angle, **cause** we know that these two angles are the same and the other two angles are the same.” | “Triangle BAM is congruent to triangle CAM. **So**, looking at side angle side **since** BM and CM are congruent, AM is congruent to itself and proved earlier BMA is congruent to CMA, **due to** side angle side, their triangles are congruent.” | “And then triangle BAM and CAM are congruent. And I'm gonna say that's because of, **so**, we know that AM, oh, nevermind. It does say that it bisects. Okay. **So**, A, so, we know that they share a congruent angle and then we know they share a line. **So**, I think it's just gonna be side angle side **because** that's really the only thing that makes sense.” |
| Evidence | Two pairs of angles are congruent and the contained sides are congruent. | Two pairs of angles are congruent. | Two pairs of sides are congruent and the contained angles are congruent. | One pair of angles is congruent and one pair of sides is congruent. |
| Warrant | Angle-Side-Angle Theorem | Angle-Side-Angle Theorem | Side-Angle-Side Theorem | Side-Angle-Side Theorem |
| Conclusion | Triangle BAM is congruent to triangle CAM | Triangle BAM is congruent to triangle CAM. | Triangle BAM is congruent to triangle CAM. | Triangle BAM is congruent to triangle CAM. |
| Evidence true? | Yes | Yes | Yes, but BM and CM have not yet been *proven* to be congruent | Yes |
| Warrant true? | Yes | Yes | Yes |  Yes |
| Valid? | Yes | No (evidence only partially meets conditions of warrant, which also requires a pair of congruent sides) | Yes | No (evidence only partially meets conditions of warrant, which also requires a second pair of congruent sides) |

Note. The correct answer is “Angle-Side-Angle Theorem.” Logical connective words in the protocols are **bolded**.

Table 6. Protocols for the fifth step of the proof in Figure 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Answer Correct? | Yes | Yes | No | No |
| Protocol | “And then it's asking BA to-- is congruent to AC. And **since** those-- I'm looking at my paper, we just use angle side angle to prove that those triangles are congruent to each other. **So, since** they're congruent, I'm going to use definition of congruent triangles to prove that those are the same.” | “AB is congruent to AC **because** they are two sides of the same triangle. But it's a definition of congruent triangles.” | “Triangle-- I mean line segment AB and AC are congruent **because** that's what they gave us.” | “AB is congruent to AC. This is **because of** definition of AB is congruent to B-- AC. Whoops. Looking at the sheet, a bisecting angle means dividing, so. Yep. It would be definition of bisecting an angle **since** it's an angle that's being bisected by the line segment.” |
| Evidence | Triangles BAM and CAM are congruent. | AB and AC are two sides of the same triangle. | Congruence of AB and AC is given. | Angle BAC is bisected by line AM. |
| Warrant | Definition of congruent triangles | Definition of congruent triangles | Given | Definition of bisecting an angle |
| Conclusion | AB and AC are congruent. | AB and AC are congruent. | AB and AC are congruent. | AB and AC are congruent. |
| Evidence true? | Yes | Yes | No, AB and AC were not given as being congruent. | Yes |
| Warrant true? | Yes | Yes | Yes |  Yes |
| Valid? | Yes | No (evidence does not meet conditions of warrant, which requires two congruent triangles) | Yes | No (evidence and warrant do not jointly imply conclusion) |

Note. The correct answer is “Definition of congruent triangles.” Logical connective words in the protocols are **bolded**.

1. More precisely, logical reasoning can establish with certainty that mathematical statements follow from axioms that are assumed to be true. [↑](#footnote-ref-2)
2. To preview our results, accuracies on the two reasoning tasks (PR: .86, GP: .62) were comparable to those in Braithwaite (2022), who did not require participants to think aloud (PR: .84, GP: .55), suggesting that thinking aloud did not substantially affect accuracy. [↑](#footnote-ref-3)
3. For the GPJT, for each proof, the periods of time from when a participant began to read a problem statement until they began to read the first step of the proof were counted as separate trials for purposes of this code. [↑](#footnote-ref-4)
4. The relation between RT-error correlation and accuracy on the GPJT was not significant when controlling for the three other concurrent skill measures, *p* = .13. [↑](#footnote-ref-5)
5. We thank an anonymous reviewer for this suggestion. [↑](#footnote-ref-6)
6. This effect remained when the other two concurrent measures that were assessed at the trial level, namely backward reference and planning/evaluating, were added as predictors, *B* = 0.39, *OR* = 1.48, *z* = 2.0, *p* = .045. [↑](#footnote-ref-7)
7. A possible concern is that participants’ warrants were usually implicit, which might seem to make it hard to judge the validity of their arguments. However, this concern does not apply in the case of true warrants, because these were *a priori* true statements, and the validity of an argument is unaffected by the inclusion or exclusion of an *a priori* true premise. The above concern is more substantial for false warrants, but is alleviated by the fact that our inferences of false warrants were mostly based on explicit statements. For example, in Table 3, protocol 4, we inferred the false warrant “If A and B are exclusive and P(C|A) = *x* and P(C|B) = *y*, then P(C|A or B) = *x* + *y*” from the explicit statement “Because it's person, not men or women, you would combine the two.” Other examples appear in the Supplementary Materials: Table E5, protocol 4 and Table E6, protocol 3. [↑](#footnote-ref-8)
8. This error seemed to reflect participants noticing that the sides whose congruence was to be proven (AB and AC), were sides of an angle known to be bisected (angle BAC), making the definition of bisecting an angle seem relevant. [↑](#footnote-ref-9)
9. By a strict definition of validity, circular reasoning (e.g., “X therefore X”) is valid, in that it is impossible for the premise to be true and the conclusion false. However, we suspect that in common usage, most people would consider circular reasoning to be invalid, so we classified it as such. [↑](#footnote-ref-10)