



Modeling the Dynamics and Autoregressive Tendencies of Metacognitive Judgment Accuracy during Complex Learning

Megan D. Wiedbusch¹ · Daryn Dever² · Tara Delgado¹ · Roger Azevedo¹

Received: 30 June 2024 / Accepted: 25 September 2025 / Published online: 21 October 2025
© International Artificial Intelligence in Education Society 2025

Abstract

Theoretical models of self-regulated learning emphasize the role of metacognitive monitoring and regulation in deploying accurate cognitive, affective, metacognitive, motivational, and social processes to improve learning outcomes. Many of these models identify metacognition as a dynamic, recursive, conditional, and contextual process. However, empirical work has largely neglected how past metacognitive judgments influence future judgments. Furthermore, prior work has primarily focused on describing what learners did during learning but has not advanced toward predicting future behavior on an individual level. Our work addresses this gap by comparing three modeling approaches to predict the accuracy of ease-of-learning (EOL) and retrospective confidence judgments (RCJs) from 55 undergraduates based on the accuracy of previous metacognitive judgments as participants learned about the human circulatory system using Metatutor-IVH a multimedia environment. We found that when fit on the individual level, the most frequently fit models were random-walk models, followed by ARIMAs, and then moving average models. Additionally, we observed consistency in the types of model fits across judgment types within individuals. When examining the accuracy in the predictions our individually fit models, ARIMA models outperformed the others in predicting RCJs. Moreover, they achieved higher accuracy than the traditional group-level models in which one model was trained on the entire sample and then fit to each individual, that our individual approach outperformed the traditional method. Finally, individuals who showed autoregressive tendencies in their EOL judgment accuracies outperformed their peers. Overall, our results make a strong argument for an individualized time-series modeling approach over traditional aggregate measures. Additionally, we found evidence that future EOL judgment accuracy may be temporally dependent on past judgments, and those that show autoregressive behaviors may be indicative of more efficient learning and metacognitive monitoring. We conclude by discussing the theoretical and practical implications for designing intelligent adaptive learning environments.

Keywords Self-regulated learning · Metacognitive judgments · Autoregressive modeling · Predictive modeling

Introduction

Self-regulated learning (SRL) is a dynamic set of cognitive, affective, metacognitive, motivational, and social processes (CAMMS) that involves the active management of one's own learning, encompassing planning, monitoring, and evaluating cognitive activities (Azevedo & Gašević, 2019; Azevedo et al., 2022; Winne, 2018). As one of the main components of self-regulated learning (SRL; Greene et al., 2024; Panadero, 2017; Winne, 2018; Winne & Azevedo, 2022), metacognition is generally defined as one's thinking about their thinking (Flavell, 1979). More specifically, metacognition can be further delineated into two primary aspects- one's knowledge and one's regulation of their cognition (Schraw & Dennison, 1994). Within the metacognitive regulatory aspects of metacognition are the processes of metacognitive monitoring. This component typically involves questioning, reflecting, drawing inferences, and making judgments about learning. These activities are dynamic and iterative, requiring learners to continuously assess and adjust their strategies to optimize learning outcomes (Greene & Azevedo, 2007; Tarricone, 2011).

There is much research that suggests that effective metacognitive monitoring can significantly enhance the efficacy of SRL (Azevedo et al., 2022; Greene et al., 2024; & Winne, 2018). This research tends to assume that monitoring is a dynamic process that can be trained such that learners become more effective and accurate in their metacognitive processes. However, this research relies primarily on aggregate and generalized measures of monitoring accuracy with vague implications about the type of instructional scaffolding that may improve one's metacognitive monitoring throughout learning. Furthermore, many of the design and scaffolding implications that can be derived from these studies are broad and somewhat ambiguous regarding support (Schraw, 2009). Specifically, these methods typically emphasize descriptive analytics of learners' behaviors, aiming to correlate these behaviors with learning outcomes such that frequency of deployment (e.g., Lim et al., 2024) or order of deployment (e.g., Azevedo et al., 2022) will inform a system's underlying scaffolding architecture. As such, this approach fails to account for the quality or accuracy of one's SRL processes.

The current study aims to address these gaps by taking an individualized dynamic modeling approach to predicting an individual's metacognitive accuracy during SRL. By focusing on the unique dynamics of learner monitoring and regulatory judgment accuracies made throughout the learning session, we can provide more precise and actionable insights into how learners monitor and regulate their cognition, thereby facilitating the design of more effective instructional interventions. While traditional metacomprehension studies have laid the groundwork for understanding metacognitive monitoring, there is a pressing need for more nuanced and individualized approaches that does not aggregate learners together. We aim to enhance our understanding through dynamic modeling techniques to enhance our understanding of

metacognitive accuracy and inform the development of adaptive instructional scaffolds that are individualized to each learner.

We begin this paper by first briefly reviewing the current state and efficacy of advanced learning technologies (ALTs) within SRL research. Within this review we highlight three potential contributing factors to the effectiveness of current ALTs that this study addresses. First, as we unpack in Sect. 3, is the misalignment between the theoretical models of SRL processes, specifically metacognitive monitoring, and the research analytical approaches commonly used. In Sect. 4 we then address the next contributing factor in which SRL scaffolds and ALTs rely primarily on aggregate versus individual student modeling. We finish our review in Sect. 5 which discusses the focus on descriptive over predictive or prescriptive analytics within ALTs. These three factors are the driving motivation for our current study which we then describe in the remainder of this article. We end on a discussion of our promising results as they relate to practical implications for future ALTs and the predictive analytics of one's metacognitive monitoring that may be used to inform future scaffolds and supports.

Literature Synthesis – ALTs Help SRL but Only so Much. Why?

Advanced learning technologies (ALTs; e.g., intelligent tutoring systems, game-based learning environments, open learner models, etc.) have significantly enhanced educational experiences and outcomes. However, their capacity to model the complex dynamics of metacognition still needs to be improved. This section examines three critical limitations of ALTs in this domain: (1) the reliance on aggregate rather than individual data for decision-making, (2) the focus on descriptive rather than predictive and prescriptive data and methods, and (3) the misalignment between theoretical constructs of metacognition and the multidimensional structure of multimodal trace data collected.

The first significant limitation of ALTs in modeling metacognition is their tendency to base decisions on aggregate data rather than individual-level information (du Boulay et al., 2023). Metacognition, the awareness and understanding of one's thought processes, is inherently personal and varies significantly among individuals (Winne, 2018). However, many ALTs rely on aggregated data from large user populations to inform their algorithms and decision-making processes coordinated between the student and expert models of typical intelligent tutoring systems (ITSs). While this approach can identify broad trends and patterns, it often fails to capture metacognitive processes' nuanced, individual nature, such as the nature, timing, quality, and relevance of the metacognitive processes (Azevedo, 2020).

For instance, while provider individualized, real-time feedback based on a learner's immediate actions, the underlying models that guide instructional sequencing and decision-making are often derived from aggregate data across populations (VanLehn, 2011, 2023). While this approach may be practical for some learners, it fails to account for individual differences in metacognitive skills, learning strategies, and self-regulation. As a result, the interventions or support provided by these systems may not be optimally tailored to each learner's unique metacognitive profile that

evolves dynamically and fluctuates over time and is contextually bound based on a complex set of learners-ITS processes that mediate learning and performance (Azevedo & Wiedbusch, 2023).

The second limitation is the predominant focus on descriptive (i.e., providing insight into what happened) rather than predictive (i.e., providing insight into what could happen) and prescriptive (i.e., providing guidance into what should be done) data and methods. Many ALTs excel at collecting and presenting descriptive data about learners' behaviors (e.g., time spent on tasks or frequency of tool use, Winne, 2017a,b), but fail to provide actionable insights into learners' metacognitive processes or offering prescriptive guidance for improvement (Conati & Lallé, 2023), which are essential for effective self-regulation (Roll & Winne, 2015). However, most ALTs lack the sophisticated algorithms to transition from descriptive to predictive and prescriptive models. It is important to note that although regression is a widely used predictive tool in metacognitive research (e.g., Ackerman & Beller, 2017; Sawyer et al., 2018; Wiedbusch et al., 2023), it is most often applied at the group level to explain variance in learning outcomes or strategy use across individuals.

The third and perhaps most challenging limitation is the need for more alignment between the theoretical constructs of metacognition (e.g., monitoring and regulation) and the multimodal trace data collected by ALTs (Azevedo, 2020). These processes are only sometimes directly observable or easily captured through digital interactions within ALTs, even with multimodal trace data (e.g., clickstream data, eye-tracking, etc.; Azevedo et al., 2018). While this multimodal data can provide insights into learners' behaviors, it must often capture the full complexity of metacognitive mechanisms and processes (Wiedbusch et al., 2023).

Moreover, theoretical models of metacognition are often high-level and abstract, proposing latent processes and mechanisms that are not directly observable. This abstraction creates challenges when trying to align these models with the fine-grained, temporally unfolding data captured by ALTs. Learners engage in complex activities with ALTs such as learning, problem solving, and reasoning processes (Molenaar et al., 2023). However, these processes evolve over multiple time scales (e.g., milliseconds to hours) which can make it difficult to map moment-by-moment behaviors onto these higher-order constructs (Ochoa, 2022). This misalignment poses significant challenges for researchers and developers. As Winne (2018) argues, there is an urgent need for more sophisticated methods of inferring metacognitive processes from observable behaviors and for developing theoretical frameworks that can bridge the gap between metacognitive constructs and digital trace data.

In conclusion, while ALTs have made significant progress in many areas of education, their ability to model the dynamics of metacognition still needs to be improved by relying on aggregate data, focusing on descriptive methods, and misaligning theory and data. Addressing these limitations will require concerted effort from researchers, developers, and educators to create more sophisticated, personalized, and theoretically grounded approaches to modeling and supporting metacognition in digital learning environments.

Theoretical Framework for Metacognitive Judgments

Theoretical models of SRL are an established set of cyclic and recursive CAMMS during learning (see Panadero, 2017 for a review of SRL models). Often within these models are a set of metacognitive monitoring tasks in which one makes judgments about their CAMMS processes throughout to determine if, when, and how to regulate their learning. This often takes the form of assessing one's progress towards goals, evaluating how well they are learning, assessing the quality of their chosen strategies and more (see Greene & Azevedo, 2009). The more accurate one is in making metacognitive judgments, the more appropriately they can then regulate their CAMMS processes which ultimately results in better performance and learning outcomes (Fleming, 2024).

For this study, we use Nelson and Narens' (1990) metamemory framework. Metamemory is defined as learners' ability to monitor and control learning and memory processes throughout a learning task. Nelson and Narens' (1990) framework establishes the relationship between monitoring and control processes where information is exchanged between the meta- and object-levels in which the meta-level is provided information from the object-level and, in turn, provides information to the object-level (Van Overschelde, 2008). This exchange of information informs the type and temporality of metacognitive judgments deployed throughout a learning task to control and change the cognitive processes that occur. The metamemory framework, and adaptations of this framework including Dunlosky & Lipko (2007), defines several metacognitive judgments (e.g., ease-of-learning, feelings of knowing, source-monitoring, confidence, judgments of learning) and control processes (e.g., processing selection, item selection, termination of study, selection of search strategy, termination of search) that a learner could enact throughout learning (Dunlosky & Bjork, 2008; Dunlosky & Lipko, 2007).

The current study focuses on two types of judgments. First, we examine ease-of-learning (EOL) judgments that learners make while acquiring information during learning. EOL judgments occur when learners initially evaluate instructional materials for levels of difficulty, prior to interacting with the material (Burkett & Azevedo, 2012; Dunlosky et al., 2013; Jemstedt et al., 2017; Mudrick et al., 2019; Wiedbusch & Azevedo, 2020). EOLs serve as the mechanism by which learners allocate attention and effort, meaning that the selection of cognitive processes and strategies as well as the effort made in deploying these strategies is driven by the EOL judgment a learner makes. For example, a learner may evaluate a topic as easy to learn, real-locating attention to a different topic that may have been initially evaluated as more difficult. Although playing a large role in controlling the cognitive processes and strategies learners enact, the study of learners' accuracy in their EOL judgments and how EOLs change over time as a function of earlier EOL judgments has been largely ignored within metacognition literature (Fleming, 2024).

The second type of judgment this study focuses on are retrospective confidence judgments (RCJs). These are judgments that occur after learning in which a learner reports how confident they are in their ability to accurately recall tested information (Stretch & Wixted, 1998; Brewer & Sampaio, 2012; Ranganath et al., 2004; Chua et al., 2006). These judgments may help learners make more effective decisions on

how to be most effective during restudy or moving forward in their learning (Papadopoulos & Obwegeser, 2021; Robey et al., 2017).

Retrospective judgments have been found to be more accurate than prospective judgments (Dougherty et al., 2018; Hines et al., 2009; Ryals et al., 2016). Some argue that this is because confidence judgments (either prospective such as EOLs or retrospective such as RCJs) are made primarily on evidence immediately made available at the moment of the decision being made (Vickers & Lee, 1998; Fleming & Diolan, 2012; Higham et al., 2009). However, more contemporary research suggests that these judgments are made as an amalgamation of evidence collected outside of the immediate judgment, and as such prospective judgments are deprived of additional information that make them less accurate (Hilgenstock et al., 2014; Siedlecka et al., 2016; Wiedbusch & Azevedo, 2020).

Measuring Metacognitive Judgment Accuracy

Metacognitive judgments can be made by a learner at local (e.g., on specific processes or questions) or global (e.g., on an entire learning session or (sub)domain) levels of performance (Seow et al., 2021; Wiedbusch & Azevedo, 2020). Schraw (2009) describes several measures of accuracy, bias, scatter, and discrimination of such judgments to evaluate performance confidence. Metacognitive accuracy is often described as the extent to which a learner's subjective confidence aligns with their objective performance (Katyal & Fleming, 2024; Schraw, 2009). While performance-based comparisons are traditional for computing judgment accuracy, this may be an oversimplification of the nuanced relationship between one's perception of their learning and their actual performance. A learner may judge a concept to be easy to learn (and thus be metacognitively accurate in terms of perceived difficulty) yet still perform poorly due to factors unrelated to monitoring accuracy, such as time pressure, anxiety, or unfamiliarity with assessment formats. Therefore, when calculating judgment accuracy, we do not assume that performance outcomes directly reflect learning quality, but rather as a measure of the alignment between a learner's subjective predictions and their actual outcomes within the constraints of a controlled multimedia learning task. Future work may benefit from incorporating models that account for external moderating factors (e.g., emotional state, time constraints, or task-switching behaviors) that may influence performance independently of metacognitive monitoring accuracy.

Typically, performance confidence judgments are made on a 1–100 unit scale, and a mean absolute accuracy score is calculated as a measure of judgment precision. The more accurate a metacognitive judgment is, the closer one's confidence is to their actual performance. As such it is not a measure of how accurate they are in their domain knowledge or task understanding. This can mean that a learner who reports 10% confidence on a question they got wrong is just as metacognitively accurate as a learner who reports 90% confidence on a question they got right. Multiple studies have found that learners who are more metacognitively accurate tend to outperform their peers in complex SRL tasks (see de Guitierrez Blume 2022 for a meta-analysis of metacognitive accuracy studies situated within SRL).

Previous metacognitive judgment research has tried to establish what may be a determinant of metacognitive accuracy (e.g., Händel et al., 2020; Koriat, 2007; Fleming & Dolan, 2012), however few studies have examined how these accuracies change over time. As Fleming and Lau (2014) reviewed, there are multiple static measures of metacognitive accuracy (e.g., ROC curves, meta-d', M-ratio). As they, and others (e.g., Azevedo, 2020) have pointed out, there is a lack of dynamic measures of metacognitive accuracy that discuss the temporal aspects of metacognitive accuracy. Desender and colleagues (2022) used binning with simulated data to calculate multiple M-ratio values per observation to identify shifts in accuracy over time. However, binning requires a theoretical or empirical justification for how to fragment time-series that are not always available.

Theoretical and Analytical Misalignment in Metacognitive Research

As we have identified above, SRL models highlight the recursive and cyclic nature of CAMMS processes (Panadero, 2017). However, the analytical methods that examine some of these processes, specifically metacognitive judgments, are often static and linear (Azevedo et al., 2010). This misalignment between theory and analytical method may be significantly impacting our ability to capture and model the dynamics of metacognitive processes during SRL, especially with ALTs. Ultimately, we hope to develop ALTs that positively affect learners into both learning domain knowledge and SRL strategies to transfer outside of the environment (Azevedo et al., 2022; Molenaar et al., 2023). However, this assumes that SRL processes change and adapt within the learning session. If our measures we use to capture and then model these processes are static measures, we may not be able to identify key moments for learners such as when they dynamically adjust their strategy usage in response to ongoing feedback and changing task demands. For example, a learner may be provided feedback that they are consistently being over confident in their judgments of how easy a task will be which is impacting how long they are examining content related to that task. After receiving this feedback, they may change how they approach making these judgments which will then impact how they approach the learning task going forward. Identifying this moment in the learning processes is more straight forward as it is driven by environmental changes.

But what happens when a learner draws a similar conclusion without any explicit feedback about their judgment accuracy? The effect may be identical, but our systems will not have available to them the moment in time this shift occurs. As such, no delimitation is made within our models and the effect may be averaged out despite the importance of such a shift. Such nuances are often lost when using static data collection and modeling methods, leading to a more superficial understanding of how metacognition functions in practice (Azevedo et al., 2005; Greene & Azevedo, 2007). This assumption is well grounded within the metacognition research which shows that while EOL judgments are typically considered task-specific (Koriat, 1997), learners often rely on generalizable cues (e.g., processing fluency and prior knowledge) to make metacognitive judgments (Murphy et al., 2022, Witherby et al., 2023). Some research argues that learners can exhibit consistent metacognitive judgments across different tasks (Miele et al., 2013), however this may be highly dependent on

task similarity (De Bruin & van Gog, 2012). Dunlosky & colleagues (2005) demonstrated that a learner's metacognitive judgment accuracy was consistent across tasks when they were in the same domain. Additionally, Hertzog et al., (2003) found that encoding fluency (i.e., the ease with which information is encoded during learning) significantly influenced judgments about future recall. These findings suggest that learners' metacognitive evaluations are not solely based on the specific content of a task but also on stable, internal cues that persist across tasks. Therefore, examining the temporal dependencies in EOL judgments could provide insights into the consistency of learners' metacognitive strategies throughout the learning session if tasks were similar enough.

Additionally, the reliance on static methods can impede the identification of causal relationships between metacognitive activities and learning outcomes. Dynamic methods, such as trace data analysis or real-time process data, enable researchers to observe the immediate effects of specific metacognitive actions, providing deeper insights into cause-and-effect relationships. Without these methods, the temporal sequence of metacognitive processes and their direct impact on learning performance remains obscured, limiting the explanatory power of the research (Winne, 2022). A possible solution to this misalignment issue might be in our modeling approaches and techniques such as time-series forecasting. However, as we will describe below, these approaches first must bend some of the traditional assumptions about modeling made within SRL research.

Student Modeling in SRL Research: Addressing the Individualization Problem in ALT Research

Most of the predictive modeling within SRL research refers to various regression approaches in which one can use a variety of predictors to model an outcome variable (or variables) using a traditional randomized controlled trial experimental design. Within these models, most commonly each data point is an individual within a sample that assumes with a large enough sample can be generalized to the population the sample was pulled from. Generalizing across learners in this manner can allow researchers to understand how a wide range of learners may engage in learning with advanced learning technologies. However, a generalized approach does not fully account for individual differences among learners (Winne, 2018). As a result, there is a loss of individual variability accounted for within student models, technologies can lack embedded personalization, diverse pathways for attaining learning outcomes are lost, and these student models can exacerbate existing educational inequities by not accounting for the specific needs of individuals. A generalized approach to student modeling can even be considered in juxtaposition to the implications for SRL.

Winne (2006, 2018, 2022) establishes the “N = me” argument which challenges traditional research paradigms to shift from focusing on group averages, aggregations, and generalizations to emphasizing the importance of considering each individual learners' differences and personalized learning approaches. The argument, situated within SRL literature, is that each learner uses different SRL strategies and processes to engage in regulation, often seeking different pathways to achieve learning out-

comes. This argument highlights that the individual learners' approach is overlooked when data is averaged across learners within a study, but could be better addressed with a shift in how data is collected and analyzed to the individual student level.

Grounding this argument are several fallacies that an individualized approach, rather than a generalized approach, would address. For example, the *ecological fallacy* posits that inferences about individuals can be accurately deduced from inferences about the group that the individual belongs to. However, as SRL research has shown, assessments about an individual from group-level averaged data can lead to incorrect assumptions about the individual and how they learn with advanced learning technologies (Fisher et al., 2018; Guglielmi & Brekke, 2017; Winne, 2022). This can have harmful effects for using big data to personalize interventions within learning technologies to scaffold individual learners. In assuming that the generalized results of aggregated data can be applied to individual learners, the learning process is considered simple rather than complex. We know from prior literature that SRL is a complex system and influenced by both internal and external components related to learning from the restrictions placed on the learning environment, to individuals' motivations, to the ecological setting in which learning takes place (Dever et al., 2022; Li et al., 2022; Marchand & Hilpert, 2024).

Additionally, at the center of the “N=me” argument is the *class reference problem*. The way in which learners deploy SRL has been linked to several cognitive conditions, including racial background, prior knowledge about a topic, motivation orientations, etc. (Winne, 2018). However, because individuals belong to multiple populations, it may be impossible to select a single group, class, or profile by which to classify an individual (Winne, 2022). As such, a single class is not able to adequately represent the variability within a single learner, especially across learning contexts (Panadero & Broadbent, 2023), leading to misleading conclusions about that learner's trajectory of SRL applications during learning tasks (Dent & Koenka, 2016). Several studies have attempted to profile learners based on their use of SRL during learning to varying degrees of success (Araka et al., 2022; Cleary et al., 2021; Dever et al., 2022; Esnaashari et al., 2023; Li et al., 2020; Wiedbusch et al., 2022). As Winne (2022) argues, instead of trying to find every possible factor that may be used to classify or group individuals, it may be better to adjust our approaches to just treat each “group” as the individual themselves.

Given the “N=me” approach, ecological fallacy, and the reference class problem, individual approaches for statistically modeling learners' SRL to better support individual differences and variability seems like an appropriate avenue for further personalization of learning environments and for better understanding how learners deploy SRL during a learning task. It comes as a double-edged sword, however, in that the instructional decisions in how scaffolds should prompt, provide feedback, and generally support individual learners no longer has the guidance that can be provided by big data and their generalizations. Because of this, it is important to consider the limitations and advantages for different modeling methods to be able to predict learners' ability and use of SRL processes as they learn about complex topics. Furthermore, the traditional regression randomized controlled trial approaches also prevent concerns of overfitting to a single sample that prevents the models from being able to be generalized to the population of interest. However, we argue that

for future scaffolds and ALTs to be adaptive to an individual, there is less concern about overfitting to an individual (who becomes our “population” of interest under the “ $N=Me$ ” approach). Instead, we should ask how generalizable these models are across contexts, and what the modeling approaches reveal about learning processes within an individual. While Winne (2018) has established this line of argumentation, there is a paucity in analytical methods to how an “ $N=Me$ ” approach could be empirically tested. We propose the following time-series modeling in which everyone is modeled independently of one another (treating them as their own “sample”). Below we briefly review this approach and how it has been used in other fields.

Using Time Series Modeling to Predict Future SRL Processes

Investigating SRL as a dynamic process, or the temporal unfolding of the CAMMS processes within SRL is not a novel concept. Many researchers have previously examined how learners traverse through the various phases of SRL through transition matrices (e.g., Dever et al., 2023; Saint et al., 2020), network analysis (e.g., Malmberg et al., 2022 2022022; Molenaar & Chiu, 2014), sequence and process mining (e.g., Li et al., 2022; Taub et al., 2022), and recursive quantification analysis (et RQA; e.g., Amon et al., 2022; Dindar et al., 2019). This research has its advantages over more descriptive research in providing more information about the sequence order and time of process deployment instead of just frequencies. For example, Dever and colleagues (2023) were able to identify some of the most common transitions between processes instead of just the frequencies of those processes themselves. This has implications for determining the “health” of the student system by examining how adaptive or repetitive a student is. RQA and network analysis also advantages in taking a systems approach to investigating SRL, especially for collaborative learning (Malmberg et al., 2022).

However, despite the strengths of these studies, their approaches are not always applicable, especially when trying to understand how processes within the same learning session impact or influence the deployment or accuracy of future processes. That is, they focus on SRL, and not the recursive or cyclic nature of any one of these processes. In addition, they prioritize log files over other measures of SRL processes such as quality or accuracy. Finally, many of these approaches (primarily Markov models, transitions matrices, and sequence mining) must assume a lack of history of prior states outside of the current one. Simply, they do not account for what a learner did outside of the current transition, or sequence of interest, in their analysis. These challenges do not negate the importance of the implications of such work, but rather limit their applications or the conclusions that can be drawn from them.

We propose the use of forecast models using everyone’s time series to train and use for predicting future metacognitive accuracies. Forecast models would allow researchers and ALT designers to account for temporal fluctuations of a learner in real-time and make predictions about an individual instead of a type, or cluster, of learner. Below, we briefly review three types of time-series forecasting models which allow for us to predict the future behavior of a time series based on the previous behavior.

Time Series Forecasting Models

Forecasting is a practice that is well established in other fields such as business (e.g., Fattah et al., 2018), meteorology (e.g., Murat et al., 2018), economics (e.g., Siami-Namini & Namin, 2018), epidemiology (Anokye et al., 2018), and many others. One of the key steps in determining if forecasting is appropriate is if the predictions can be more accurate than chance by capturing the genuine patterns and relationships of historical data (Panagiotelis & Hyndman, 2021). While other modeling approaches, such as regression, seek to discover what factors may influence the outcome of interest, simple time series forecasting only use the historical past of the variable to make predictions. Decomposition methods can also help uncover seasonal or trend patterns within time series. For this study, we only introduce one variable to be modeled and predicted- metacognitive judgment accuracy. However, we discuss the implications of more sophisticated time series forecasting techniques in our discussion. Additionally, we treat everyone as their own sample to examine if taking a “N=me” approach to our analysis improves modeling accuracy without the addition of any information on a learner’s individual differences. While many time series forecasting methods exist, we limit our study to three which would help direct future studies and approaches. These methods, their advantages, limitations, and assumptions are described below.

Moving Average

Moving average (MA) models are a type of forecast that assume the next observation in a time series is influenced by the linear combination of the past white noise error. These models examine the general trend of the data by eliminating some of the randomness of the time series through smoothing and decomposition. Instead of modeling specific past values, these models focus on broader trends by smoothing and reducing randomness through decomposition. These models are less concerned with the small fluctuations within a time series and more for the overall trend-cycle. That is, these models are helpful when we are interested in modeling the overall direction in a process rather than making precise predictions for each individual point. Simply the average method in which the forecast of all future values (test data) is equal to the average of the historical data ($y_1, \dots, y_T$). We can write our forecasts as

$$\hat{y}_{T+h|T} = \bar{y} = (y_1 + \dots + y_T)/T \quad (1)$$

Within our study, we can imagine these models concerned about predicting the general change in one’s metacognitive judgment accuracy over the learning session and less about the individual predictions of a single future observation. If we were to see most individuals fitting to a moving average model, we would assume that learners are making more general or holistic reflections about their past metacognitive accuracy when making their next observation rather than any single judgment.

Random Walk

Random walk models, also known as a naïve forecast, take a different approach. They assume the best prediction for the next time point is simply the most recent value without regard for earlier observations or average behavior. These models are often metaphorically described as a drunkard's walk where, while attempting to walk in a straight line, either takes a random step left or right. Each step depends only on where they just were, not their full path. While they appear similar to the walking average models, they are anchored to the last observed value instead of the average. This means that these models become less and less accurate the further out they try to make a forecast. We can write our forecasts as

$$\hat{y}_{T+h|T} = y_T \quad (2)$$

In our study, this model represents a type of learner whose metacognitive judgments are anchored entirely in the most recent experience. That is, when deciding how accurate they currently are, they rely solely on how accurate they felt just one trial ago, without incorporating any earlier feedback. If learners' judgments follow a random walk, it would suggest a short-memory strategy, focused on moment-to-moment adjustments rather than cumulative patterns.

ARIMA

The third modeling approach this study explores is the most sophisticated – autoregressive integrated moving average (ARIMA) models. These models are considered a general class of models that combine three components. First, is the autoregressive (AR) portion which uses past values to predict current ones. Next, is the integrated (I) portion which involves differencing the series to make it stationary. That is, it transforms the data such that the overall level and variability of the data stays consistent over time without any trends or “seasonal” swings. Finally, is the moving average portion (MA) which models the error term as a linear combination of past forecast errors. Simply, an ARIMA forecast model creates a linear equation in which the predictors include both lagged observations and lagged forecast errors.

As they are a class of models in which model identification is data-driven, we have no a-prior assumptions about which of these components are necessary for any given time-series. This means an ARIMA model could be simplified to both random walk (ARIMA(0,1,0)) and moving average (ARIMA(0,0, q)) models. Therefore, for our study, a model was considered an ARIMA model fit if it contained an autoregressive component, as described in Sect. 7.4.3.

The identification of an ARIMA model is defined by a set of three parameters (p, d, q) in which p is the order of the autoregressive part (i.e., the number of autoregressive terms or past values), d is the degree of the number of times the time series has been differenced to achieve stationarity, and q is the order of the moving average part (i.e., the number of lagged forecast errors). For instance, an ARIMA(1,1,1) model would first difference the original series once to remove trends ($d=1$) and then

predict the next value using both the immediately previous difference values ($p=1$) and the previous forecast error ($q=1$). We can write our ARIMA forecast as

$$y'_t = c + \varphi_1 y'_{t-1} + \cdots + \varphi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad \hat{y}_{T+h|T} = y_T \quad (3)$$

Where ϵ_t is white noise, y'_t is the time series after being difference d times. Although the differencing parameter d does not appear in the equation directly, it is reflected in the transformed data being modeled. That is, Eq. 3 is applied after differencing has made the series stationary. Within our study, these models would assume that the current judgment a learner is making is influenced by not just the immediately previous one, but by patterns in prior judgments (the AR component), systematic shifts or trends in performance (addressed through differencing), and the average or accumulated past forecasting errors (the MA component). For example, a learner's confidence on a current task may be shaped by both specific prior evaluations and their accumulated deviation from accuracy—capturing both their reflective tendencies and noise in judgment.

Current Study

The learning sciences field is in large agreement that individualized support within ALTs will ultimately provide beneficial learning outcomes, especially compared to a one-size-fits-all pedagogical approach (Azevedo & Gašević, 2019). However, the current deployment of scaffolds and supports within ALTs still rely on aggregate models of “similar” learners. As Winne (2018) highlights and we have just expanded upon, this approach is susceptible to the ecological fallacy and reference class problems which may be contributing to the less than large effect sizes we are currently seeing despite advancing technological capabilities. Furthermore, our methodological and analytical approaches are not always in alignment with the theoretical models of SRL and (meta)cognitive processes they aim to measure. Specifically, the field highlights the recursive and cyclic nature of metacognition but continues to rely on aggregate metrics over entire learning sessions. Finally, the field has also not progressed past a primarily descriptive stage in our work into more predictive (and ultimately prescriptive) stage. When prediction is discussed, it is at the student level (based on similar populations) and not within an individual (e.g., predicting the next behavior, next strategy use, next judgment made, etc.).

We begin to address these three gaps by exploring metacognitive judgment accuracy prediction by purposefully fitting models to individuals learning within MetaTutor-IVH, a multimedia learning environment which prompts learners to provide multiple metacognitive judgments.

MetaTutor IVH

MetaTutor-IVH (Intelligent Virtual Human) is a multimedia learning environment that uses student's ease of learning [EOL] judgments, content evaluations [CEs], and

retrospective confidence judgments [RCJs] to scaffold learners via a non-verbal pedagogical agent. This linearly structured environment prompts learners' metacognitive judgment as they learn while the agent also provides its' own judgments throughout the task. MetaTutor IVH's content covers nine human physiological systems (e.g., cardiovascular, musculoskeletal, nervous, etc.).

Previous work with MetaTutor-IVH has supported previous metamemory research, finding that the more accurate metacognitive judgments were, the better one's performance and learning was (Mudrick et al., 2017; Wiedbusch & Azevedo, 2020; Wiedbusch et al., 2023). Specifically, these studies found that the accuracy of one's EOLs (Mudrick et al., 2017) and content evaluations (Wiedbusch & Azevedo, 2020; Wiedbusch et al., 2023), were predictive of learning. Other research with MetaTutor-IVH has also suggested that metacognitive judgments should not be considered as independent phenomenon but as an intricate and interconnected process (Wiedbusch & Azevedo, 2020; Wiedbusch et al., 2022). For example, one study found that individuals could be clustered across metacognitive judgments and these clusters significantly differed in their overall performance (Wiedbusch et al., 2022). This study found that learners who tend to have high EOL judgment accuracy additionally have high RCJ judgment accuracy and made some theoretical implications about local and global metacognitive activity in complex learning environments such as MetaTutor-IVH.

Research Questions and Hypotheses

Specifically, in this study, we examined three modeling approaches that predicted future metacognitive judgment accuracy indices from previous metacognitive judgment accuracies. That is, we examined the predictive nature of EOLs on future EOLs, and RCJs on future RCJs (prior to and post a written justification). In this paper, we discuss the predictive accuracy of each approach, the advantages and disadvantages, and the recommended use cases based on our results. Our approaches include a random-walk forecast model (based on the last judgment accuracy), moving average forecast model (based on the historical average of past judgment accuracy), and an auto-regressive model (fit using a variation of the Hyndman-Khandakar algorithm).

Given limited previous empirical work but more extensive theoretical assumptions, we raise the following research questions and hypotheses.

RQ1: Which modeling approach is most frequently the best fit across participants for each metacognitive judgment?

Given the limited empirical research but more extensive theoretical assumptions, we predict that most participants' best fitting model will be a moving average forecast model for each type of metacognitive judgment. This model is the simplest, and typical metacognitive judgment research (e.g., Jemstedt et al., 2017; Mudrick et al., 2019; Wiedbusch & Azevedo, 2020, 2023) tends to focus on the average accuracy over an entire learning session.

RQ2: Is the model of best fit consistent across all metacognitive judgments for each participant?

Given the research that has examined the accuracy of different types of metacognitive judgments within the same learning task (e.g., Mudrick et al., 2017; Siedlecka et al., 2016), we predict that we will see the best fit model for prospective (i.e., ease-of-learning) judgments differ from the retrospective judgments (i.e., retrospective confidence judgments both pre and post judgment) for each participant. However, we predict that the best fit modeling approach will be consistent within the retrospective judgments for each participant.

RQ3: Which modeling approach is most accurate across participants for each metacognitive judgment?

Given that we are unable to report any previous studies that use this approach, we make our hypothesis based entirely off theoretical assumptions about the cyclic and recursive nature of metacognition (Katyal & Fleming, 2024; Peters, 2022; Winne, 2018). As such, we predict that for participants whose best fitting model was an ARIMA model, they will have the most accurate forecasting. We predict this will hold true for all metacognitive judgment types.

RQ4: Is there a significant difference in the average accuracy between using the same modeling approach for all participants or using their best fitting model?

As we have established above, aggregating individuals opens ourselves up to several fallacies that may impact our ability to offer individualized scaffolding and support (Winne, 2018). As such, we predict that the average accuracy when we fit an individual model to each individual learner will have higher overall accuracy than fitting the same model to all individuals.

RQ5: Are there differences in learning outcomes between best fitting model groups for each metacognitive judgment?

Given previous research that indicates more quality deployment of metacognitive processes during learning is associated with higher learning outcomes and the theoretical assumptions that metacognitive processes are cyclic and recursive in nature, we hypothesize that participants with an autoregressive model as their best fitting model will have higher learning gains than participants with other best fitting model types. We predict this will hold true for all metacognitive judgment types.

Methods

Participants

Fifty-five undergraduate students from a large northeastern university were given an opportunity to learn about various human body systems using MetaTutor-IVH, a multimedia learning environment. The participants gender distribution is 65%

female, 22% male, and 13% that did not report their gender. Their ages ranged from 18 to 30 ($M=20.64$, $SD=2.68$). 56% of students reported having taken or being currently enrolled in a biology course. Additionally, 75% were pursuing bachelor's degrees with majors from STEM fields. All participants were compensated \$10/hr up to \$30 for their participation. Human research approval was received prior to recruitment and data collection.

Experimental Procedure

All participants were required to first complete a demographic questionnaire, an 18-item multiple-choice biology content pretest, and a series of self-report measures assessing emotions, perceptions, and motivation (i.e., Achievement Emotions Questionnaire [AEQ (Pekrun et al., 2011)], Emotion Regulation Questionnaire [ERQ (Gross & John, 2003)], Perceived Affect Utility Scale [PAUSE (Chow & Berenbaum, 2012)]) prior to interacting with MetaTutor-IVH. None of the self-reports were utilized in this study, so for brevity we do not expand more. Upon completion of the self-reports, participants were then instrumented and calibrated with wireless shimmer 3+ electrodermal activity bracelet, eye tracker, and affect recognition software. During this experiment, multimodal data, including eye tracking, facial expressions, EDA, and log files were collected. For this study, only log files were used that detailed the human-computer interaction between the participant and the system. Participants were then instructed that they were about to learn about different body systems by reading text, inspecting diagrams, and answering questions about the content all within the MetaTutor-IVH environment. Afterward, participants then completed the 18 trials of MetaTutor-IVH before another round of the same battery of self-report measures. All participants were then thanked for their participation and provided with monetary compensation.

Experimental Design

The data analyzed in the present study were drawn from a larger experiment originally designed to investigate the effects of pedagogical agent affect on learning and student emotion within MetaTutor-IVH. That study employed an 18 trial, $3 \times 3 \times 2$ within-subjects design in which trials are defined by content relevancy (fully relevant, text somewhat relevant, diagram somewhat relevant), agent facial expressions (congruent, incongruent, neutral), and question type (function, malfunction). While the original purpose of the study emphasized affective and emotional processes, the present paper focuses on a secondary analysis of the same dataset to examine learners' metacognitive judgements within this environment. This reuse is motivated by the richness of the dataset and the large difference in the questions being addressed about metacognition and affect. Given the cost and complexity of running large-scale multimodal education studies, while the data in this current study has been previously used (see Mudrick et al., 2017; Wiedbusch & Azevedo, 2020, 2022), this analysis addresses new questions and research focus.

Content relevancy (fully relevant, text somewhat relevant, diagram somewhat relevant) refers to how well the source of the relevant content aligns with the question.

Fully relevant content can be found in the text and diagram. Text somewhat relevant means the relevant content can be found in the diagram, while the text is only somewhat relevant material. Diagram somewhat relevant means the relevant content can be found in the text, while the diagram is only somewhat relevant material. There is always enough information to answer every multiple-choice question, however, the source of the relevant material was experimentally manipulated.

Agent facial expression (congruent, incongruent, neutral) refers to emotions presented by the tutor during its' content evaluation (described below). When the agent expresses joy (smiling), it insinuated that the agent felt all the content was relevant. Expressions of confusion (a head tilt and brow furrow) insinuated the agent believing only some of the content was relevant. Finally, a neutral expression (no change from baseline), did not make any insinuation about the content relevancy.

Lastly, question types are divided into malfunction versus function. Each body system was given two trials asking about the body when it was functioning correctly (e.g., describing how much oxygen we inhale versus exhale) versus when the body system was incorrectly functioning (e.g., the effects of cystic fibrosis on the respiratory system).

Within each of the 18 trials described above, there are seven major steps (see Fig. 1). First, the learning environment presents a question to learners that they will have to answer. They are then prompted to provide an EOL metacognitive judgment by answering the question *“How easy do you think it will be to learn the information needed to answer this question?”* at the center of the screen. Responses are collected on a 1-unit incremental scale from 0 to 100.

The third step begins after this judgment was provided, and the learner is presented with a content page. This page consists of three paragraphs along the left hand of the screen, one diagram in the center, and a pedagogical agent at the top right-hand corner. After a ten-second delay the learner is prompted to provide two content-evaluation (CE) metacognitive judgments by answering the questions, *“Do you feel the content (text/diagram) on this page is relevant to the question?”* After the response from the learner is received, the tutor then provides its' own content evaluation metacognitive judgment about the relevancy of the content. This judgment is provided as a non-verbal facial expression of joy, confusion, or neutral and would be given after 10 s.

Once a learner is finished reviewing the content slide, they then proceed to answer the multiple-choice question. After submitting their choice, they were asked to provide a retrospective confidence judgment (RCJ1) by answering *“How confident are you that the answer you provided is correct?”* on a 1-unit incremental scale from 0 to 100. The last steps first required the learner to provide a written justification for their answer and then provide another retrospective confidence judgment (RCJ2). At the trial's end, the learner is guided to the subsequent trial.

Data Analysis Approach

Each participant ($N=55$) provided multiple metacognitive judgments per trial ($n=18$). For this study, we examined 1 prescriptive judgment (EOL) taken prior to any content and 2 retrospective judgments (RCJ1 and RCJ2) taken post multiple-

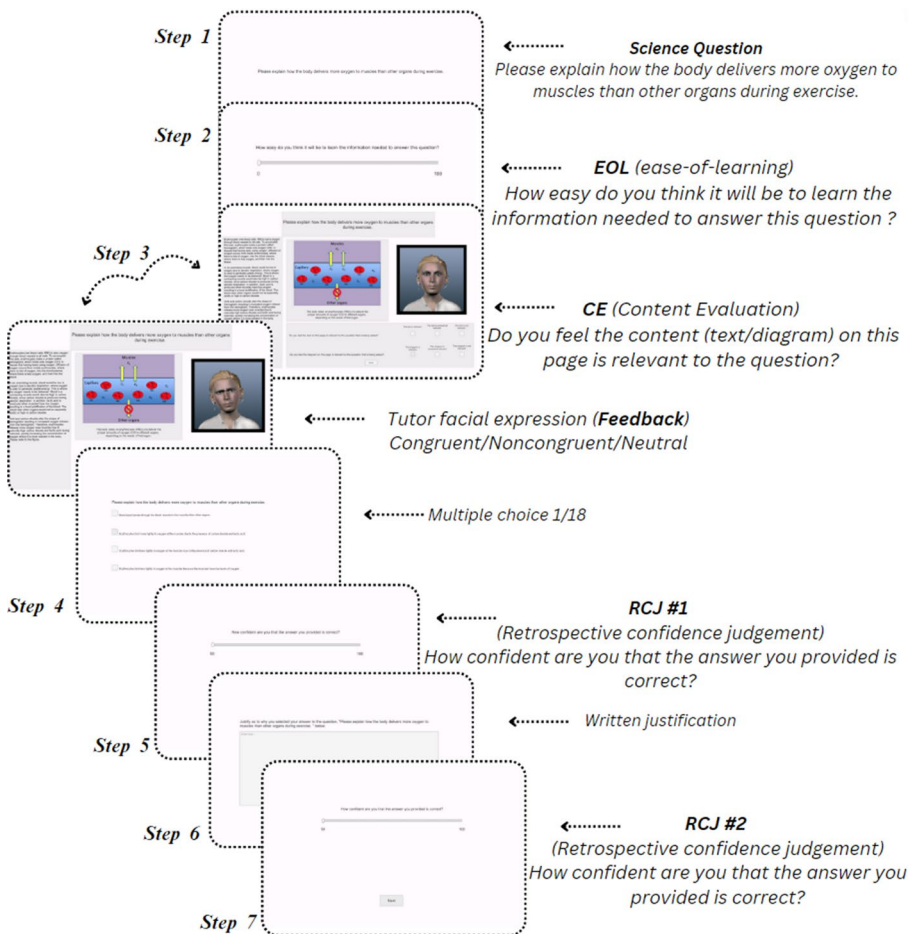


Fig. 1 MetaTutor-IVH trial breakdown

choice question. RCJ1 was prompted prior to a written justification for the learner's answer and RCJ2 was prompted post his justification. For RQ1-RQ3, we examine each of these judgments independently of one another but follow the same modeling procedure (see 3.6) using metacognitive judgment accuracy (see 7.4.1). Learning outcomes were operationalized by calculating normalized learning gains (see 7.4.2).

Metacognitive Judgment Accuracies

Judgment accuracy is the degree to which a learner's perception of their learning is aligned to an objective measure of the learning performance. For example, EOL accuracy is how well a learners' estimate of how easy a question will be to answer (without previewing any instructional material) matches their actual performance on answering the question.

For each trial, we calculated the Brier Score (BS) for each participant. The BS (see Eq. 4) is a measure of the squared difference between the confidence rating c and decision accuracy o where $o=0$ for incorrect answers on the multiple-choice question and $o=1$ for correct answers on the multiple-choice question. BS scores are error scores, and as such during interpretation, the lower the BS value, the more accurate a judgment is. The BS is nearly identical to Schraw's (2009) Absolute Accuracy Index which is more commonly used within metamemory research. However, this index is calculated over a series of judgments while a BS can be calculated over a single judgment. This is necessary if we hope to make predictions about future individual judgments.

$$BS = \sum_i (o_i - c_i)^2 \quad (4)$$

Normalized Learning Gains

We operationalized learning outcomes by a learner's normalized learning gain (NLG; see Eq. 5; Marx & Cummings, 2007) which measures how much a participant learns accounting for prior knowledge. Specifically, a change score (c) in domain knowledge is calculated using the equation:

$$c = \begin{cases} \frac{post-pre}{100-pre} & post > pre \\ drop & post = pre = 100 \text{ or } 0 \\ 0 & post = pre \\ \frac{post-pre}{pre} & post < pre \end{cases} \quad (5)$$

Modeling Approach

Before fitting any models, we split each set of 18 judgment accuracies (e.g., EOL) into a training set (or historical data; $n=14$) and a testing set (or future data; $n=4$), following the widely accepted 80/20 Pareto principle (Dunford et al., 2014). We then categorized each learner by their best fitting modeling approach by automatically fit each training data set to an ARIMA(p,d,q) model using a variation of the Hyndman-Khandakar algorithm using R's *forecast* package (Hyndman & Khandakar, 2008; Hyndman et al., 2023). We used the Akaike information criterion corrected (AICc) to determine best fit.

In ARIMA modeling, we do not have an a-priori forecast model before the model identification occurs. This identification determined by the parameters (p,d,q) in which p is the order of the autoregressive part, d is the degree of the first differencing involved, and q is the order of the moving average part. As such, some ARIMA models can be "simplified" to AR(p) which indicates weighted moving average over past observations. I(d) which indicates a linear or polynomial trend but no autocorrelation, also known as a random-walk model. MA(q) indicates a weighted moving average over past errors. ARMA(p,q) indicates an auto-regressive moving average,

or $\text{IMA}(d,q)$ which indicates an integrated moving average, often related to an exponential smoothing model.

For the purposes of this study, we classified each participant's best fitting model as either (1) a random-walk forecast model (based on the last judgment accuracy,) (2) a moving average forecast model (based on the historical average of past judgment accuracy), or (3) an auto-regressive model. That is, models that returned $p > 0$ were classified as an auto-regressive model (regardless of d and q parameter values). Models that returned $p = 0$ and $q > 0$ were categorized as moving average models (regardless of d parameter values). Models that returned $\text{ARIMA}(0,0,0)$ which is a white noise model, we fit a random-walk model as a participant's best fitting model. Additionally, models that returned as $\text{ARIMA}(0,1,0)$ were categorized as random walk models.

Model Accuracy Metrics

To determine how accurate a model was at predicting the test data after being fit on the training data, we calculated the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Mean Absolute Scaled Error (MASE). As these are all error scores, the lower the value, the more accurate the prediction model.

Results

Preliminary Results

We first examined the distribution and central tendencies of each metacognitive judgment accuracy's Brier Scores for all participants. All distributions were deemed normally distributed as skewness was within ± 2 and kurtosis were within ± 7 (Byrne, 2013; Hair et al., 2010) for each participant's 18 trials for each judgment (EOL, RCJ1, and RCJ2). Table 1 reports the descriptive statistics for the means of each judgment for each participant.

RQ1: Which Modeling Approach is Most Frequently the Best Fit Across Participants for Each Metacognitive Judgment?

We automatically fit each training data set to an $\text{ARIMA}(p,d,q)$ model using a variation of the Hyndman-Khandakar algorithm using R's *forecast* package (see Sect. 7.4.3

Table 1 Descriptive statistics for the mean metacognitive judgment Brier scores across participants

Metacognitive Judgment	Min	Max	Median	Mean	Std. Dev	Skewness	Kurtosis
EOL	0.23	0.52	0.31	0.34	0.07	0.79	-0.29
RCJ1	0.25	0.59	0.42	0.42	0.08	0.08	-0.53
RCJ2	0.23	0.63	0.40	0.42	0.10	0.33	-0.82

*The assumption of normality was determined on the participant level, but we only report the mean descriptives for brevity here

for more details on our modeling approach). We used the AICc to determine best fit. The distribution of best fitting models and their AICc are reported in Table 2.

Across all three judgment types, the most frequent model fit to a participant was a random walk model, followed by an ARIMA model, and then a moving average model (see Fig. 2 for exemplar forecast visualizations for EOL judgments).

To examine if there were significant differences in AICc between the Random Walk models and the ARIMA models, we conducted a Welch's t-test (due to unequal sample sizes). No significant difference was found ($t(23.77) = -1.42, p = 0.17$) suggesting that these two approaches had approximately equivalent fits. Given the small sample size of the moving average model group ($N = 3$), we excluded them from

Table 2 Model fit frequency and AICcs for all metacognitive judgments

	Metacognitive Judgment					
	EOL		RCJ1		RCJ2	
Model Type	N	$M(SD)_{AICc}$	N	$M(SD)_{AICc}$	N	$M(SD)_{AICc}$
Random Walk	38	8.08 (12.20)	32	16.57 (8.76)	28	16.50 (8.54)
Moving Average	3	9.28 (1.49)	8	15.50 (5.07)	6	15.01 (5.47)
ARIMA(p,d,q)	14	2.59 (12.60)	15	10.31 (10.70)	21	7.670 (13.00)

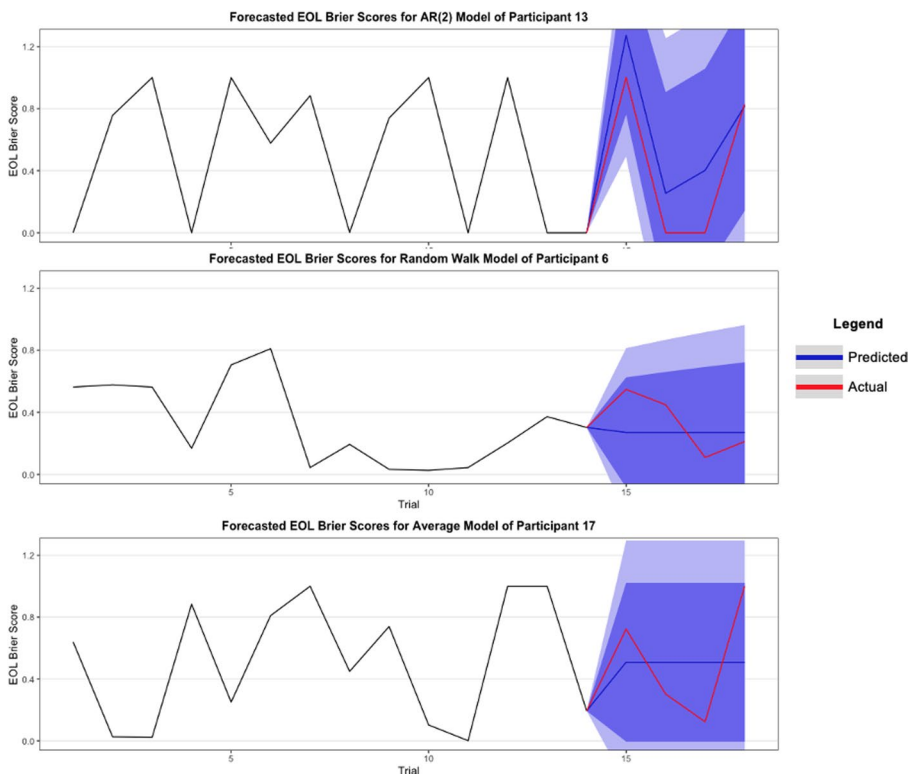


Fig. 2 Forecasted EOL Brier score exemplars for each modeling approach

comparison of the model fit frequencies. This suggests that when a model was fit a participant, no one approach was a better fit on average.

Table 3 reports the ARIMA model parameter frequencies for each judgment for participants who best fitting model was an ARIMA. We can see from this table that for all metacognitive judgments, the order of first autoregressive part was roughly uniformly distributed across parameter values ranging from 1 to 4. Additionally, the degree of the first differencing and order of the moving parts were primarily 0 for most participants.

RQ2: Is the Model of Best Fit Consistent across all Metacognitive Judgments for each participant?

We wanted to see if individuals would have consistent best model types across metacognitive judgments. Figure 3 maps the frequencies of the model types. From this, we can see that the most frequent model combination was all Random Walk models. Additionally, we see a clustering of the next most frequent model combinations (4–6 frequency range, back right across top two levels of Fig. 3) in which all three models were a combination of Random Walk and ARIMAs (with a slight preference for ARIMAs for RCJ1s). These findings are consistent with RQ1 and suggest that there is some regularity in the modeling approach within individuals.

RQ3: Which Modeling Approach is Most Accurate across Participants for each Metacognitive Judgment?

Next, we examined the accuracy for each participant's modeling approach in forecasting the last four trials. As previously mentioned, all models were fit only on the first 14 trials, and then tested for accuracy on the last 4 (roughly 20%). We report all model accuracy metrics (see Sect. 7.4.4) in Table 4 for each metacognitive judgment across all model types.

For each metacognitive judgment, we report four accuracy metrics (see Sect. 7.4.4). To determine if there were any significant differences in any of the accuracy metrics within each metacognitive judgment across the model types, we conducted a multivariate Kruskal-Wallis test (the non-parametric equivalent of a MANOVA, chosen due to violation of the normality assumption; He et al., 2017). For the first test on EOL judgments, we only compared the Random Walk models to the ARIMA models due to low sample size in the moving average model group. We found no significant difference in the accuracy metrics across the model types for EOL models ($H(4)=6.78$, $p=0.15$). Next, we examined all three model groups for both RCJs. We found a significant difference in the accuracy metrics across model types for both RCJ1 ($H(8)=15.61$, $p=0.05$) and RCJ2 ($H(8)=19.35$, $p=0.01$) metacognitive judgments. Using the Multivariate Scheffé-like technique outlined in Katz and McSweeney (1980) for post-hoc analysis, however, we failed to find any significant differences between model types for either judgment across all accuracy metrics. This is likely due to the conservative nature of this specific post-hoc comparison approach which utilizes the ranks of the accuracy metrics while also accounting for repeated measures.

Parameter	Definition		
Parameter	Definition	EOL	RCJ1
			RCJ2

[illegible]

Best Fit Model Types Across Metacognitive Judgements

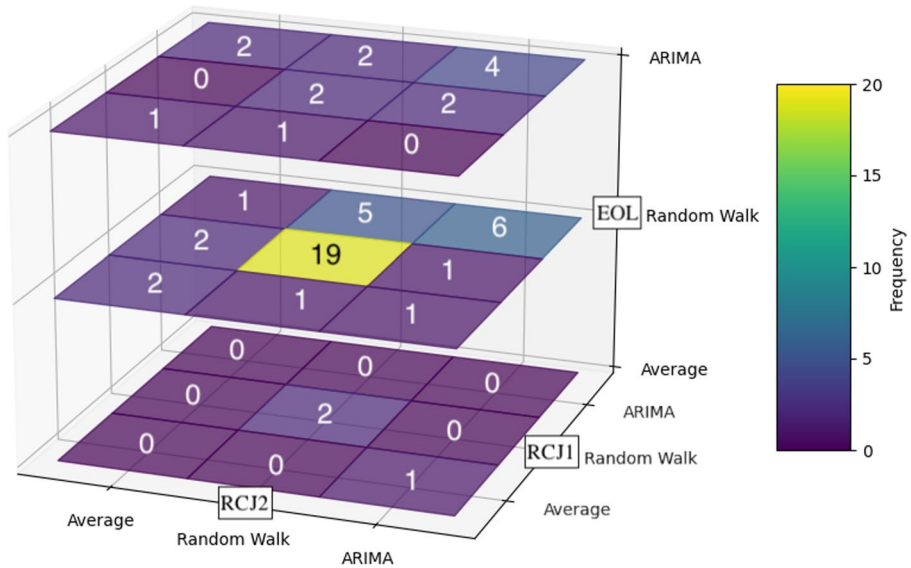


Fig. 3 Frequencies of best fit models for EOLs (y-axis), RCJ1s (z-axis), and RCJ2 (x-axis)

Table 4 Model accuracy metrics means and standard deviations by model type for all metacognitive judgments

Meta-cognitive Judgment	Model Type	Model Accuracy Metrics			
		$M(SD)_{MAE}$	$M(SD)_{RMAE}$	$M(SD)_{MAPE}$	$M(SD)_{MASE}$
EOL	Random Walk	0.20 (0.10)	0.23 (0.11)	340.34 (780.63)	1.36 (2.80)
	Moving Average	0.24 (0.11)	0.29 (0.11)	4122.72 (7020.40)	0.94 (0.28)
	ARIMA(p,d,q)	0.18 (0.10)	0.20 (0.10)	629.30 (1212.04)	0.89 (0.54)
RCJ1	Random Walk	0.29 (0.14)	0.31 (0.13)	18995.38 (58707.22)	1.17 (1.04)
	Average	0.36 (0.15)	0.40 (0.15)	266.99 (400.41)	0.90 (0.41)
	ARIMA(p,d,q)	0.35 (0.13)	0.39 (0.14)	1778.99 (4570.81)	1.66 (2.92)
RCJ2	Random Walk	0.29 (0.17)	0.31 (0.16)	5883.87 (17976.36)	3.25 (8.48)
	Moving Average	0.39 (0.23)	0.46 (0.24)	49807.04 (119650.30)	0.83 (0.52)
	ARIMA(p,d,q)	0.26 (0.12)	0.30 (0.14)	12101.83 (33164.42)	0.91 (0.85)

*depicts significant difference between model levels within each judgment type at $p < 0.05$

In sum, we found that for EOL judgments, the accuracy metrics across models were not significantly different. Additionally, while we did find significant differences in the model accuracy metrics for the RCJ metacognitive judgment models, we are unable to determine which groups significantly differed from one another. We take these findings in conjunction to suggest that when a participant's best model fit was an ARIMA model, they were more accurate in predicting the final 20% of trials' RCJ metacognitive judgment accuracy.

RQ4: Is there a Significant Difference in the Average Accuracy Between Using the Same Modeling Approach for All Participants or Using their Best Fitting Model?

Given that individualizing the modeling approach does not appear to favor one approach over another, we then wanted to understand if this individualized approach was more accurate than if we chose to aggregate all participants into one modeling approach (hereby referred to as the “traditional” approach). To make this comparison, our first step was to use our same modeling approach from above on the first 14 trials of the average metacognitive judgment accuracy of all participants to train one model. We then used this model to determine the accuracy for predicting everyone’s final four trials. Table 5 reports the average accuracy of each analytical approach below using our accuracy metrics for each of the three metacognitive judgment types.

To determine if there were any significant differences in any of the accuracy metrics within each metacognitive judgment across the two analytical approaches, we conducted a multivariate Kruskal-Wallis test (the non-parametric equivalent of a MANOVA, chosen due to violation of the normality assumption; He et al., 2017). We found no significant difference the accuracy metrics between the analytical approaches ($H(4)=3.00$, $p=0.56$). However, we did find significant differences between the analytical approaches for both RCJ1 ($H(4)=19.50$, $p<0.001$) and RCJ2 ($H(4)=10.23$, $p=0.04$). Using the Multivariate Scheffé-like technique outlined in Katz and McSweeney (1980), we then conducted post-hoc analysis for RCJ1 and RCJ2 judgments across analytical approaches to determine which of the accuracy metrics was contributing to the significant differences between the analytical approaches. We failed to identify the significant difference for the RCJ2 judgments. This is likely due to the conservative nature of this specific post-hoc comparison approach which utilizes the ranks of the accuracy metrics while also accounting for repeated measures. However, we identified that the MAE and RMAE metrics were significantly different between our two approaches for RCJ1 (see Table 6). In sum, we found that using an individualized approach was more accurate in predicting RCJ accuracy than using a model built off aggregate data.

RQ5: Are There Differences in Learning Outcomes Between Best Fitting Model Groups for each Metacognitive Judgment?

Next, we wanted to assess if learners who showed different types of behaviors based on their best fit modeling approach differentiated in their learning outcomes. Specifically, we wanted to see if those who showed any indication of autoregressive tendencies outperformed their peers. To do so, we conducted three Kruskal Wallis comparison tests (non-parametric ANOVA equivalent), one for each metacognitive judgment due to unequal sample size and normality assumption violations. Table 7 reports the results of these tests which indicated there was a significant difference between modeling approaches for EOL models ($\chi^2(2)=7.89$, $p=0.02$).

To determine which modeling approaches significantly differed among each other, we then conducted a series of Dunn Tests for post hoc comparisons for the EOL judgments (see Table 7). We found a significant difference between the Random Walk models ($M_{\text{NLG Rank}} = 23.99$) and the ARIMA(p,d,q) models ($M_{\text{NLG Rank}} = 37.14$;

Table 5 Model accuracy metrics means and standard deviations by analytical approach type for all meta-cognitive judgments

Meta-cognitive Judgment	Analytical Approach	Model Accuracy Metrics			
		$M(SD)_{MAE}$	$M(SD)_{RMAE}$	$M(SD)_{MAPE}$	$M(SD)_{MASE}$
EOL	Traditional	0.20 (0.10)	0.23 (0.11)	625.24 (1859.63)	1.21 (2.34)
	Individualized	0.21 (0.10)	0.23 (0.10)	672.14 (1747.20)	1.19 (1.65)
RCJ1	Traditional	0.31 (0.14)	0.35 (0.14)	12343.14 (47055.73)	1.27 (1.71)
	Individualized	0.22 (0.09)	0.25 (0.09)	828.15 (2141.58)	1.29 (1.30)
RCJ2	Traditional	0.30 (0.16)	0.32 (0.17)	13766.41 (47142.16)	2.09 (6.14)
	Individualized	0.22 (0.09)	0.25 (0.09)	817.90 (2116.06)	1.28 (1.31)

Table 6 Multivariate Scheffé-like technique for post hoc comparisons for model accuracy metrics between analytical approach for RCJ1 judgments

Accuracy Metric	95% CI	Mean Rank Deviation	Significant Conclusion
MAE	[-20.23, -1.49]	-10.86	*
RMAE	[-21.05, -2.31]	-11.68	*
MAPE	[-13.85, 4.89]	-4.48	ns
MASE	[-5.63, 13.10]	3.74	ns
$\chi^2(4) = 31.19$, Scheffé Critical Value = 9.49, $p < 0.01$			

*Note, * depicts significance at $\alpha = 0.95$

Table 7 Kruskal Wallis comparison of model group NLGs for each metacognitive judgment

Meta-cognitive Judgment	Model Type	N	M_{NLG} Rank	SD_{NLG} Rank	Kruskal-Wallis $\chi^2(2)$	Adjusted p
EOL	Random Walk	38	23.99	15.80	7.89	0.02*
	Moving Average	3	33.50	17.20		
	ARIMA (p, d, q)	14	37.14	12.38		
RCJ1	Random Walk	32	27.94	16.37	0.11	0.95
	Moving Average	8	26.56	12.79		
	ARIMA (p, d, q)	15	28.90	17.67		
RCJ2	Random Walk	28	27.95	17.21	0.19	0.91
	Moving Average	6	30.58	12.42		
	ARIMA (p, d, q)	21	27.33	15.85		

*Note, * depicts p significant at $p < 0.05$. All ps have been adjusted using the Holm-Bonferroni method for multiple comparisons

Table 8 Dunn test for post hoc comparisons for ranked normalized learning gains based on best fitting model types for EOL judgments

Comparison	Group 1 M(SD) _{NLG Rank}	Group 2 M(SD) _{NLG Rank}	Group 1 N	Group 2 N	z	Ad-just-ed P
Random Walk vs. Moving Average	23.99 (15.80)	33.50 (17.20)	38	3	0.32	0.64
Random Walk vs. ARIMA(p,d,q)	23.99 (15.80)	37.14 (12.38)	38	14	2.74	0.02*
Moving Average vs. ARIMA(p,d,q)	33.50 (17.20)	37.14 (12.38)	3	14	0.41	0.68

*Note, * depicts p significant at $p < 0.05$. All ps have been adjusted using the Holm-Bonferroni method for multiple comparisons

$z=0.41$, $p_{adj} = 0.02^*$). In sum, we found that for individuals whose EOL accuracy showed signs of autoregressive tendencies outperformed their peers.

Discussion

The overall goal of this study was to examine an approach to predicting the accuracy of one's metacognitive judgments based on the accuracy of the previous judgments. To do so we employed three modeling approaches to help reveal potential differences among learner deployment of accurate metacognitive judgments and predict their future accuracy. We also examined the differences in learning gains between modeling approaches.

The first of these approaches, a moving average method, assumes that most of the variability in one's accuracy of EOLs is white noise (i.e., random). This approach is the most complementary to the contemporary analysis of metacognitive accuracy and bias which looks at the overall accuracy of a series of judgments. However, this approach also then assumes that previous judgment accuracy has no influence over future judgment accuracy. Or in other words, we can learn and improve our metacognitive accuracy within a learning session. This assumption would not bode well for the design of scaffolds and supports in advanced learning technologies (e.g., intelligent tutoring systems) specifically aimed at supporting and improving metacognitive processes of learners. Unsurprisingly, we found that the lowest number of individuals were fit to a moving average model across all judgment types.

Our second approach, a random walk model suggests that the next judgment's accuracy takes a random step away from its previous value, but these steps are independent of one another. These models appear similar to the moving average method models but are anchored to last observed value instead of the average of the historical data. Within our context, this would suggest that the accuracy of metacognitive judgments is only really accounting for the most recent accuracy which would mean future scaffolds shouldn't require large datasets before making predictions to deploy support. Most individuals, across all judgment types, were fit to random walk models. In addition to the theoretical implications of what information informs learners of their next judgment, we also take this finding to have actionable design implications. Random walks are not computationally demanding to implement, further supporting our argument to have systems model individuals without requiring the storage of large data sets or accounting for large moving windows within data sets.

Furthermore, our traditional approach of aggregating individuals to train one model (suggesting a random walk model) and then fit to an individual was outperformed by our individualized modeling approach (in which the majority were also random walk models). This suggests that even if the final modeling approach is the same, only using an individual's data may be more beneficial in predicting an individual's next behavior or (meta)cognitive process over aggregating to the mean.

Our final set of models were autoregressive models that forecast future metacognitive accuracy based on historical data. These models would suggest that previous metacognitive experiences are influencing future metacognitive judgments. This is something that many theoretical models of metacognition assume but that current analytical approaches largely ignore. The modeling approaches chosen in conjunction with our results suggest that even without feedback on the accuracy of their metacognition, about 30% of learners had autoregressive tendencies (past values influence future ones). Furthermore, these learners tended to outperform their peers as shown in RQ5. That is, learners who appear to reflect on their past metacognitive judgments and adjust were better self-regulators and outperformed their peers. It is important to note that while learners whose EOL judgments were best modeled with autoregressive patterns (ARIMA) showed higher learning gains, this does not necessarily indicate that their judgments were more metacognitively accurate. Rather, the autoregressive fit suggests that these learners' current judgments were more systematically influenced by prior judgments. This may reflect a more consistent or strategic monitoring approach but does not alone imply improved alignment between perceived and actual performance. Therefore, the association between autoregressive patterns and learning outcomes should be interpreted as a potential indicator of structured self-monitoring, not as direct evidence of enhanced metacognitive calibration.

We argue that this effect would occur for more learners and be more pronounced if they were provided explicit feedback on their accuracy throughout the learning session to help them make these evaluations and adjustments more quickly. As further evidence of this interpretation, it is important to note that this finding was only for the EOL judgments as we failed to find evidence in differences in learning gains for either of the RCJs. Learners may have more salient and temporally close information (i.e., their multiple choice and written response they just provided) by which to judge their RCJ on compared to EOLs in which a new trial has started allowing for a more holistic reflection. This pattern aligns with previous work showing that momentary subjective experiences, such as immediate feedback or emotion reactions, may serve as dominant cues for item level judgments (Ackerman, 2019; Koriati, 1997). This in turn may have made RCJs less reliant on past experiences and more influenced by immediate task features.

Importantly, although we use performance-aligned accuracy metrics such as the Brier Score, we recognize that these do not fully capture the complexity of metacognitive monitoring, especially for judgments like EOLs that occur prior to task engagement. In future work, we aim to integrate real-time affective and behavioral data (e.g., time-on-task, gaze, or self-reported effort) to create more holistic models of judgment accuracy that go beyond performance metrics alone.

While our findings provide important insight into the dynamic structure of metacognitive judgment accuracy, it is important to consider the role of our highly con-

trolled task design in shaping these patterns. The 18 learning trials in MetaTutor-IVH were embedded in a $3 \times 3 \times 2$ within-subjects design that systematically manipulated content relevancy, agent facial expressions, and question type. This structure was essential for maintaining internal validity and ensuring a broad sampling of metacognitive judgments across varied learning conditions. However, we acknowledge that this level of experimental control may have influenced both the frequency and stability of judgment patterns observed. For instance, systematic variation in agent feedback or content relevancy may have differentially affected learners' metacognitive calibration, thereby shaping the model fits. Consequently, the relative prevalence of random walk or ARIMA models in our study may not directly translate to more naturalistic or less-structured learning environments. Future research should explore how these modeling approaches perform in open-ended or student-directed learning contexts, where learners self-regulate task sequences and durations, and where external scaffolds are less prescriptive. Such work will help to determine the extent to which these time-series forecasting models remain robust across ecologically valid learning environments. It is also important to note the limitation of our sample size ($N=55$, each with 18 repeated trials). While repeated-measures designs increase statistical power by capturing within-person variance, the modest number of participants constrains the generalizability (across both context and learners) of our findings. Replication with larger and more diverse samples will be necessary to confirm whether the observed modeling patterns, particularly the prevalence of random walk and autoregressive fits, hold across broader populations and learning contexts.

Autoregressive models could inform future development of scaffolds by helping learners understand when they are over- and under-confident in their judgments and how this may impact their overall learning goals. Future research on the impact of supplying learners with their metacognitive accuracy as feedback is needed to understand how one can learn from their judgments over time. Additionally, more research is needed to explore what the influence of context (e.g., domain topic, prior knowledge, individual differences, scaffolding) may have on modeling choices.

Conclusion

There have been many repeated calls for individualized and adaptive learner models in learning technologies and predictive learning analytics, however, conventional approaches rely almost exclusively on generalized or aggregate models to explain variance in metacognitive judgment accuracy. Additionally, these approaches tend to use static measures of metacognitive accuracy, misaligned with the theoretical grounding of these processes within SRL. Both have contributed to the lack of predictive modeling (compared to descriptive) within the field to help make prescriptive suggestions about one's learning.

Using the “ $N=Me$ ” argumentation from Winne (2018), we have proposed a unique methodological approach to forecasting one's metacognitive accuracy. Specifically, we have examined various patterns across approaches across individuals. That is, instead of generalizing a method across participants, we are exploring individualized modeling methods by accepting the assumption that we may “overfit” to an

individual (which is only a concern if we wanted to use their model to generalize to another learner). However, more systematic testing is needed to test if overfitting to an individual within a certain context is also occurring. This would lessen the viability of this study's approach as the generalizability would be too heavily contextualized. However, our study seems to suggest this may not be the case. Our study expands current analytical approaches by incorporating predictive modeling techniques, revealing potential autoregressive tendencies in repeated metacognitive judgments made during complex learning. More work is needed to explore if this tendency holds when examining different judgments and the predictive nature of one judgment on another.

Overall, our work highlights the possibility of treating everyone as their own population by which we fit models to. This in turn opens our field up to other analytical approaches, such as time-series forecasting, seen in other fields. If we can reliably use these models within new ALTs, we will then be able move towards more prescriptive scaffolds. For example, we can imagine a system that asks a learner to self-test while also providing a metacognitive judgment about how confidence they are in their answer at various locations throughout studying about the human circulatory system. Using data from similar self-testing questions when the learner studied microbiology, the system identified an autoregressive tendency in which they consistently reference the last two quizzes every time they engage in a self-test task. However, the system knows that the first quiz they took would be more informative based on content. It might then deploy a scaffold to this particular learner to reflect on this quiz and section of their studying instead of the most recent two sections. Alternatively, if a student's model (categorized as a moving average model) suggests that they reflect on all their previous quizzes, the system would instead provide a scaffold based on over or under confidence of the task in general. That is, instead of referencing a particular section of learning, it reminds the student that they tend to be overconfident in how well they are actually studying and may wish to revisit that particular section before moving on. The deployment of both scaffolds is informed by the accuracies of their metacognitive judgment but are uniquely designed based on the type of model that best fit to their individual data and that model's subsequent inference about their SRL processes. These decisions could be further informed and supplemented by using unobtrusive measures of metacognitive judgments such as eye-tracking or think-alouds as well.

While more work is needed to understand how these model fits may change over time, or how assumptions of generalizability within an individual are met across contexts, our study has contributed to the future design and development of perspective scaffolds that are individualized to a learner by only being modeled to that individual. Furthermore, our work has begun to explore possible autoregressive tendencies within metacognitive judgments to help empirically support some of the theoretical assumptions of metacognitive monitoring within SRL. The findings underscore the necessity of moving beyond traditional aggregate models to embrace approaches that account for individual differences and temporal dynamics. Such advancements hold promise for improving the design and implementation of ALTs, ultimately supporting learners in becoming more effective and autonomous in their learning.

Acknowledgements The authors would like to thank the members from the SMART Lab at UCF and the IntelliMedia Group at NCSU for their assistance.

Author Contributions Conceptualization, methodology, analysis were performed by Megan Wiedbusch. Drafting and refinement of the manuscript were performed by all authors.

Funding Disclosure.

The research presented in this paper is supported by funding from the National Science Foundation, DRL 1431532, Roger Azevedo.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

- Ackerman, R., & Beller, Y. (2017). Shared and distinct cue utilization for metacognitive judgments during reasoning and memorisation. *Thinking & Reasoning*, 23(4), 376–408.
- Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psihologijske teme*, 28(1), 1–20.
- Amon, M. J., Mattingly, S., Necaise, A., Mark, G., Chawla, N., Dey, A., & D'mello, S. (2022). Flexibility versus routineness in multimodal health indicators: A sensor-based longitudinal in situ study of information workers. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(3), 1–27. <https://doi.org/10.1145/3514259>
- Anokye, R., Acheampong, E., Owusu, I., Isaac Obeng, E., & Lin, Y. (2018). Time series analysis of malaria in kumasi: Using ARIMA models to forecast future incidence. *Cogent Social Sciences*, 4(1). <https://doi.org/10.1080/23311886.2018.1461544>
- Araka, E., Oboko, R., Maina, E., & Gitonga, R. (2022). Using educational data mining techniques to identify profiles in self-regulated learning: An empirical evaluation. *The International Review Of Research In Open And Distributed Learning*, 23(1), 131–162. <https://doi.org/10.19173/irrodl.v22i4.5401>
- Azevedo, R. (2020). Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition And Learning*, 15, 91–98. <https://doi.org/10.1007/s11409-020-09231-x>
- Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, 96, 207–210. <https://doi.org/10.1016/j.chb.2019.03.025>
- Azevedo, R., & Wiedbusch, M. (2023). Theories of metacognition and pedagogy applied in AIED systems. In du B. Boulay, A. Mitrovic, and, & K. Yacef (Eds.), *Handbook of artificial intelligence in education* (pp. 141–173). Edward Elgar Publishing.
- Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science*, 33(5–6), 381–412. <https://doi.org/10.1007/s11251-005-1273-8>
- Azevedo, R., Moos, D. C., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist*, 45(4), 210–223. <https://doi.org/10.1080/00461520.2010.515934>
- Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 254–270). Routledge.

- Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., & Cerezo, R. (2022). Lessons learned and future directions of metatutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. *Frontiers in Psychology*, 13, Article 813632. <https://doi.org/10.3389/fpsyg.2022.813632>
- Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, 67(1), 59–77. <https://doi.org/10.1016/j.jml.2012.04.002>
- Burkett, C., & Azevedo, R. (2012). The effect of multimedia discrepancies on metacognitive judgments. *Computers in Human Behavior*, 28(4), 1276–1285. <https://doi.org/10.1016/j.chb.2012.02.011>
- Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.
- Chow, P. I., & Berenbaum, H. (2012). Perceived utility of emotion: The structure and construct validity of the perceived affect utility scale in a cross-ethnic sample. *Cultural Diversity & Ethnic Minority Psychology*, 18(1), 55–63. <https://doi.org/10.1037/a0026711>
- Chua, E. F., Schacter, D. L., Rand-Giovannetti, E., & Sperling, R. A. (2006). Understanding metamemory: Neural correlates of the cognitive process and subjective level of confidence in recognition memory. *NeuroImage*, 29(4), 1150–1160. <https://doi.org/10.1016/j.neuroimage.2005.09.058>
- Cleary, T. J., Slemp, J., & Pawlo, E. R. (2021). Linking student self-regulated learning profiles to achievement and engagement in mathematics. *Psychology in the Schools*, 58(3), 443–457. <https://doi.org/10.1002/pits.22456>
- Conati, C., & Lallé, S. (2023). Student modeling in open-ended learning environments. In B. du Boulay, A. Mitrovic, & K. Yacef (Eds.), *Handbook of Artificial Intelligence in Education* (pp. 170–183). Edward Elgar.
- De Bruin, A. B., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, 22(4), 245–252.
- Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, 28(3), 425–474. <https://doi.org/10.1007/s10648-015-9320-8>
- Desender, K., Vermeylen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications*, 13(1), Article 4208. <https://doi.org/10.1038/s41467-022-31727-0>
- Dever, D. A., Amon, M. J., Vrzakova, H., Wiedbusch, M. D., Cloude, E. B., & Azevedo, R. (2022). Capturing sequences of learners' self-regulatory interactions with instructional material during game-based learning using auto-recurrence quantification analysis. *Frontiers in Psychology*, 13, Article 813677. <https://doi.org/10.3389/fpsyg.2022.813677>
- Dever, D. A., Sonnenfeld, N. A., Wiedbusch, M. D., Schmorrow, S. G., Amon, M. J., & Azevedo, R. (2023). A complex systems approach to analyzing pedagogical agents' scaffolding of self-regulated learning within an intelligent tutoring system. *Metacognition And Learning*, 18(3), 659–691. <https://doi.org/10.1007/s11409-023-09346-x>
- Dindar, M., Alikhani, I., Malmberg, J., Järvelä, S., & Seppänen, T. (2019). Examining shared monitoring in collaborative learning: A case of a recurrence quantification analysis approach. *Computers in Human Behavior*, 100, 335–344. <https://doi.org/10.1016/j.chb.2019.03.004>
- Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory & Cognition*, 46(4), 558–565. <https://doi.org/10.3758/s13421-018-0791-y>
- du Boulay, B., Mitrovic, A., & Yacef, K. (2023). *Handbook of artificial intelligence in education*. Edward Elgar.
- Dunford, R., Su, Q., & Tamang, E. (2014). The Pareto principle. *The Plymouth Student Scientist*, 7(1), 140–148.
- Dunlosky, J., & Bjork, R. A. (2008). The integrated nature of metamemory and memory. *Handbook of Metamemory and Memory* (pp 11–28).
- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, 16(4), 228–232. <https://doi.org/10.1111/j.1467-8721.2007.00509.x>
- Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, 52(4), 551–565.

- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>
- Esnaashari, S., Gardner, L. A., Arthanari, T. S., & Rehm, M. (2023). Unfolding self-regulated learning profiles of students: A longitudinal study. *Journal of Computer Assisted Learning*, 39(4), 1116–1131. <https://doi.org/10.1111/jcal.12830>
- Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*. <https://doi.org/10.1177/1847979018808673>
- Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27), E6106–E6115.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1), 241–268.
- Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1338–1349.
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, Article 443.
- Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review Of Educational Research*, 77(3), 334–372.
- Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary educational psychology*, 34(1), 18–29.
- Greene, J. A., Bernacki, M. L., & Hadwin, A. F. (2024). Self-regulation. *Handbook of Educational Psychology* (pp. 314–334).
- Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2), 348.
- Guglielmi, R. S., & Brekke, N. (2017). A framework for understanding cross-national and cross-ethnic gaps in math and science achievement: The case of the United States. *Comparative Education Review*, 61(1), 000–000.
- de Gutierrez Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal Of Educational Psychology*, 114(4), 681.
- Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. In *Multivariate data analysis* (pp. 785–785).
- Händel, M., de Bruin, A. B., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning*, 15(1), 51–75.
- He, F., Mazumdar, S., Tang, G., Bhatia, T., Anderson, S. J., Dew, M. A., & Reynolds, C. F., III. (2017). Non-parametric MANOVA approaches for non-normal multivariate outcomes with missing values. *Communications in Statistics - Theory and Methods*, 46(14), 7188–7200.
- Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(1), 22.
- Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1), 57.
- Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd better think twice: Post-decision perceptual confidence. *NeuroImage*, 99, 323–331.
- Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology And Aging*, 24(2), 462.
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, (3), 1–22. <https://doi.org/10.18637/jss.v027.i03>
- Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2023). Forecast: Forecasting functions for time series and linear models. R Package Version 8.21.1. <https://pkg.robjhyndman.com/forecast/>

- Jemstedt, A., Kubik, V., & Jönsson, F. U. (2017). What moderates the accuracy of ease of learning judgments? *Metacognition And Learning*, 12, 337–355.
- Katyal, S., & Fleming, S. M. (2024). The future of metacognition research: Balancing construct breadth with measurement rigor. *Cortex*, 171, 223–234.
- Katz, B. M., & McSweeney, M. (1980). A multivariate Kruskal-Wallis test with post hoc procedures. *Multivariate Behavioral Research*, 15(3), 281–297.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349.
- Koriat, A., & Helstrup, T. (2007). Metacognitive aspects of memory. In S. Magnussen & T. Helstrup (Eds.) *Everyday memory* (pp. 251–274). Psychology Press.
- Li, S., Chen, G., Xing, W., Zheng, J., & Xie, C. (2020). Longitudinal clustering of students' self-regulated learning behaviors in engineering design. *Computers & Education*, 153, 103899.
- Li, S., Zheng, J., & Lajoie, S. P. (2022). Temporal structures and sequential patterns of self-regulated learning behaviors in problem solving with an intelligent tutoring system. *Educational Technology & Society*, 25(4), 1–14.
- Lim, L., Bannert, M., van der Graaf, J., Fan, Y., Rakovic, M., Singh, S., & Gašević, D. (2024). How do students learn with real-time personalized scaffolds? *British Journal of Educational Technology*, 55(4), 1309–1327.
- Malmberg, J., Saqr, M., Järvenoja, H., Haataja, E., Pijeira-Díaz, H. J., & Järvelä, S. (2022). Modeling the complex interplay between monitoring events for regulated learning with psychological networks. In M. Giannakos, D. Spikol, D. Di Mitri, K. Sharma, X. Ochoa, & R. Hammad (Eds.), *The multimodal learning analytics handbook* (pp. 79–104). Springer International Publishing.
- Marchand, G. C., & Hilpert, J. C. (2024). Contributions of complex systems approaches, perspectives, models, and methods in educational psychology. In P. Schutz & K. Muis (Eds.) *Handbook of educational psychology* (pp. 139–161). Routledge.
- Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, 75(1), 87–91.
- Miele, D. B., Son, L. K., & Metcalfe, J. (2013). Children's naive theories of intelligence influence their metacognitive judgments. *Child development*, 84(6), 1879–1886.
- Molenaar, I., & Chiu, M. M. (2014). Dissecting sequences of regulation and cognition: Statistical discourse analysis of primary school children's collaborative learning. *Metacognition And Learning*, 9, 137–160.
- Molenaar, I., de Mooij, S., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139, Article 107540.
- Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Computers in Human Behavior*, 96, 223–234.
- Mudrick, N. V., Taub, M., Azevedo, R., Rowe, J., & Lester, J. (2017). Toward affect-sensitive virtual human tutors: The influence of facial expressions on learning and emotion. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 184–189). IEEE.
- Murat, M., Malinowska, I., Gos, M., & Krzyszczak, J. (2018). Forecasting daily meteorological time series using ARIMA and regression models. *International Agrophysics*, 32(2), 253–264. <https://doi.org/10.1515/intag-2017-0007>
- Murphy, D. H., Huckins, S. C., Rhodes, M. G., & Castel, A. D. (2022). The effect of perceptual processing fluency and value on metacognition and remembering. *Psychonomic Bulletin & Review*, 29(3), 910–921.
- Nelson, T. O., & Narens, L. (1990). Why investigate metacognition. *Metacognition: Knowing about Knowing*, 13, 1–25.
- Ochoa, X. (2022). Multimodal learning analytics: Rationale, process, examples, and direction. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *The handbook of learning analytics* (2nd ed., pp. 54–65). Society for Learning Analytics Research (SoLAR).
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, Article 422.
- Panadero, E., & Broadbent, J. (2023). Self-regulated learning in higher education: Theoretical perspectives and practical applications. *Educational Psychology Review*, 35, 123–145. <https://doi.org/10.1007/s10648-023-09645-1>

- Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, 37(1), 343–359.
- Papadopoulos, P. M., Obwegeser, N., & Weinberger, A. (2021). Concurrent and retrospective metacognitive judgments as feedback in audience response systems: Impact on performance and self-assessment accuracy. *Computers and Education Open*, 2, Article 100046.
- Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology*, 36(1), 36–48.
- Peters, M. A. (2022). Towards characterizing the canonical computations generating phenomenal experience. *Neuroscience & Biobehavioral Reviews*, 142, 104903.
- Ranganath, C., Cohen, M. X., Dam, C., & D'Esposito, M. (2004). Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. *Journal of Neuroscience*, 24(16), 3917–3925.
- Robey, A. M., Dougherty, M. R., & Buttafio, D. R. (2017). Making retrospective confidence judgments improves learners' ability to decide what not to study. *Psychological Science*, 28(11), 1683–1693.
- Roll, I., & Winne, P. H. (2015). Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics*, 2, 7–12.
- Ryals, A. J., Rogers, L. M., Gross, E. Z., Polnaszek, K. L., & Voss, J. L. (2016). Associative recognition memory awareness improved by theta-burst stimulation of frontopolar cortex. *Cerebral Cortex*, 26(3), 1200–1210.
- Saint, J., Whitelock-Wainwright, A., Gašević, D., & Pardo, A. (2020). Trace-SRL: A framework for analysis of microlevel processes of self-regulated learning from trace data. *IEEE Transactions on Learning Technologies*, 13(4), 861–877.
- Sawyer, R., Mudrick, N. V., Azevedo, R., & Lester, J. (2018). Impact of learner-centered affective dynamics on metacognitive judgments and performance in advanced learning technologies. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19* (pp. 312–316). Springer International Publishing.
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition And Learning*, 4, 33–45.
- Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4), 460–475.
- Seow, T. X., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological Psychiatry*, 90(7), 436–446.
- Siami-Namini, S., & Namin, A. S. (2018). *Forecasting economics and financial time series: ARIMA vs. LSTM*. arXiv preprint arXiv:1803.06386.
- Siedlecka, M., Paulewicz, B., & Wierchoń, M. (2016). But i was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, 7, Article 218.
- Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 24(6), 1379.
- Tarricone, P. (2011). *A taxonomy of metacognition*. Academic Press.
- Taub, M., Banzon, A. M., Zhang, T., & Chen, Z. (2022). Tracking changes in students' online self-regulated learning behaviors and achievement goals using trace clustering and process mining. *Frontiers in Psychology*, 13, Article 813514.
- Van Overschelde, J. P. (2008). Metacognition: Knowing about knowing. *Handbook of Metamemory and Memory*, 47, 71.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46, 197–221.
- VanLehn, K. (2023). Evaluations with aided systems. In du B. Boulay, A. Mitrovic, & K. Yacef (Eds.), *Handbook of artificial intelligence in education* (pp. 505–523). Edward Elgar.
- Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics Psychology and Life Sciences*, 2, 169–194.
- Wiedbusch, M. D., & Azevedo, R. (2020). Modeling metacomprehension monitoring accuracy with eye gaze on informational content in a multimedia learning environment. In *ACM symposium on eye tracking research and applications* (pp. 1–9).

- Wiedbusch, M. D., Sonnenfeld, N., Dever, D., & Azevedo, R. (2022). Clustering learner's metacognitive judgment accuracy and bias to explore learning with AIED Systems. In Proceedings of the International Conference on Artificial Intelligence in Education (pp. 402–413). Cham: Springer International Publishing.
- Wiedbusch, M., Lester, J., & Azevedo, R. (2023). A multi-level growth modeling approach to measuring learner attention with metacognitive pedagogical agents. *Metacognition & Learning* 18, 465–494. <https://doi.org/10.1007/s11409-023-09336-z>
- Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist*, 41(1), 5–17.
- Winne, P. H. (2017a). Learning analytics for self-regulated learning. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 241–249). Society for Learning Analytics Research.
- Winne, P. H. (2017b). Theories of self-regulated learning viewed from models of information processing. In D. H. Schunk & B. J. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 36–48). Routledge.
- Winne, P. H. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology*, 88(1), 9–20.
- Winne, P. H. (2022). Modeling self-regulated learning as learners doing learning science: How trace data and learning analytics help develop skills for self-regulated learning. *Metacognition Learning*, 17, 773–791. <https://doi.org/10.1007/s11409-022-09305-y>
- Winne, P., & Azevedo, R. (2022). Metacognition and self-regulated learning. *The Cambridge Handbook of the Learning Sciences* (pp. 93–113).
- Witherby, A. E., Carpenter, S. K., & Smith, A. M. (2023). Exploring the relationship between prior knowledge and metacognitive monitoring accuracy. *Metacognition And Learning*, 18(2), 591–621.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Megan D. Wiedbusch¹  · Daryn Dever²  · Tara Delgado¹  · Roger Azevedo¹ 

✉ Megan D. Wiedbusch
megan.wiedbusch@ucf.edu

¹ School of Modeling, Simulation, and Training, University of Central Florida, Orlando, FL 32826, USA

² College of Education, University of Florida, Gainesville 32611, United States of America