What's Next? Sequence Length and Impossible Loops in State Transition Measurement

Nigel Bosch
University of Illinois
Urbana–Champaign
pnb@illinois.edu

Luc Paquette
University of Illinois
Urbana–Champaign
lpaq@illinois.edu

Transition metrics, which quantify the propensity for one event to follow another, are often utilized to study sequential patterns of behaviors, emotions, actions, and other states. However, little is known about the conditions in which application of transition metrics is appropriate. We report on two experiments in which we simulated sequences of states to explore the properties of common transition metrics (conditional probability, D'Mello's *L*, lag sequential analysis, and Yule's *Q*) where results should be null (i.e., random sequences). In experiment 1, we found that transition metrics produced statistically significant results with non-null effect sizes (e.g., $Q > 0.2$) when sequences of states were short. In experiment 2, we explored situations where consecutively repeated states (i.e., loops, or self-transitions) are impossible – e.g., in digital learning environments where actions such as hint requests cannot be made twice in a row. We found that impossible loops affected all transition metrics (e.g., $Q = .646$). Based on simulations, we recommend sequences of length 50 or more for transition metric analyses. Our software for calculating transition metrics and running simulated experiments is publicly available.

Keywords: sequential analysis, transition metrics, simulated data

1. INTRODUCTION

Suppose two statistics students are learning about the normal distribution. Both students start off engaged, but over time one student becomes frustrated by the difficulty of the learning task, eventually disengages, and becomes bored. Meanwhile, the other student finds the task too easy and also becomes bored. Both students are experiencing boredom – a highly undesirable emotion for learning (Baker et al., 2010) – but for different reasons. Situations like these have motivated a growing body of research that analyzes the sequence of behavioral, emotional, cognitive, or other states that students experience (Baker et al., 2007; Bosch & D'Mello, 2017; D'Mello & Graesser, 2012; McQuiggan et al., 2010, 2008). A key requirement for these studies is a method to measure the tendency for one state (e.g., an emotion) to follow another. These methods consist of various *transition metrics*, which quantify probability, likelihood, or other measures of how often one state transitions to another.

When used effectively, transition metrics allow researchers to examine students' behaviors from state to state over time, thereby revealing new insights into learning behaviors that might otherwise remain hidden (Chen et al., 2018; Knight et al., 2017). For the example statistics students above, we might find that the first student's tendency to transition through frustration to boredom indicates very different learning behaviors than the second student's tendency to transition directly from engagement to boredom. However, some key properties of transition metrics are as yet not well understood, especially how sequence length relates to results and how differences in the type of sequence may influence results (e.g., non-repeatable vs. repeatable action sequences in a computerized learning environment). We focus on these two issues in the current paper, uncovering some common situations where transition metrics may produce flawed results and comparing these results across metrics.

1.1. ISSUES OF SEQUENCE LENGTH

Educational research examines state sequences of various lengths, ranging from less than ten to thousands (Andres et al., 2019; Ocumpaugh et al., 2017), depending on the type of data and how the data were collected. However, little is known about how sequence length relates to transition metric results. Shorter sequences might yield more noise, for example, if the standard deviation of a statistic calculated on these data is higher, but it is unclear whether such noise might lead to systematic biases in results. Moreover, it is unclear how long sequences should be to minimize possible biases. Hence, we explore this issue for several different transition metrics in this paper to provide guidance on what sequence lengths researchers should collect when studying state transitions and what kinds of problems related to sequence length they should anticipate and monitor.

1.2. ISSUES OF STATE TYPE

Sequences consist of states including emotions, behaviors, actions in a user interface, or any other discrete characterization of educational experiences that might lead to insight into learning. The type of state and manner of data collection may have important implications for results from transition metric analyses, however. For example, if researchers examine transitions between different regions of a user interface, such as tabs (Biswas et al., 2016) or activities (Bosch & D'Mello, 2017), it is impossible for sequences to include two consecutive occurrences of the same state. Similarly, certain student actions may not be allowed to repeat, such as pressing a hint button twice in a row without taking any actions based on the first hint or pressing a "submit answer" button twice in a row (for two consecutive exercises) without entering an answer for the second exercise. In a graph of transitions such as Figure 1, these consecutively repeating states are represented by arrows from a state to itself and are referred to as state persistence, self-transitions, or *loops* (in graph theory terminology). We refer to situations where loops cannot arise (e.g., consecutive hint requests) as sequences with *impossible loops*.

In the above examples, states and the transitions between them can be perfectly observed via log file analysis. Certain state types cannot be perfectly observed, however, and must be obtained via methods such as self-reports (Larson & Csikszentmihalyi, 1983) or periodic observation (Ocumpaugh et al., 2015). For example, attentional states are often observed via sampling methods (e.g., self-report, classroom observation). In such cases, the transitions between states are not directly observed; rather, they are inferred from points where an observed state differs from the previously observed state. In these cases, the same state may be observed several times in a row. For example, suppose a student pays attention for five minutes; if researchers prompt them to self-report attention once per minute, their sequence data will include five consecutive "paying attention" reports before transitioning to a new state. However, these reports represent only one incident of attention, not five. Increasing the frequency of prompts will lengthen the sequence of consecutive identical reports, despite the underlying phenomenon not changing at all. Thus, researchers sometimes remove or ignore consecutive occurrences of the same state to focus on the transitions only (Bakeman & Quera, 1995; Bosch & D'Mello, 2017; D'Mello & Graesser, 2012; Karumbaiah et al., 2019). As a consequence, the same situation may arise as in analyses of user interfaces, where a state cannot be consecutively repeated. We examine one such scenario in this paper using data from a computer-based learning environment (see section 3.3) where events are recorded each time a student completes an activity. We explore transitions between activities, where loops are possible, or exercise IDs, where loops are not possible and must be removed from the sequence (they occur in the data because exercise ID is recorded repeatedly for each activity within an exercise). Such situations are relatively common in educational data, where one type of state sequence is nested within another and requires loop removal to obtain the actual transitions.

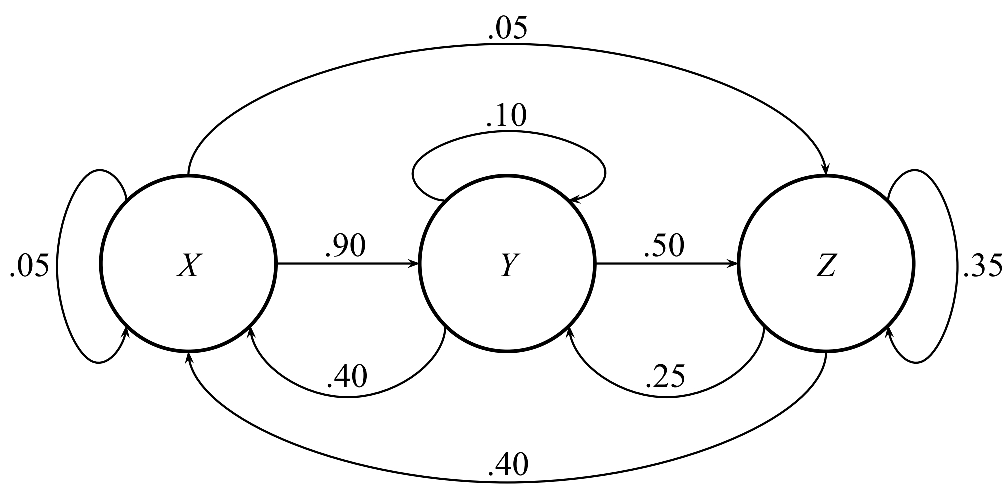

Figure 1: Example transition probabilities between three states in a sequence. Transitions form a Markov chain model, where the probability of transitioning to any state is conditioned only on the one previous state in the sequence.

1.3. CONTRIBUTION AND NOVELTY

In this paper, we describe results from a series of simulations that illustrate the properties of common transition metrics, then compare them to two different types of sequence data from a computer-based learning environment. We explore relevant cases that occur in real-world scenarios, including short sequences and sequences where loops are impossible, and show how transition metric results are affected by these scenarios. We expect that short sequences will exacerbate issues caused by state type and random noise and thus seek to discover how long sequences must be to avoid large adverse effects on results.

We make recommendations for proper usage of transition metrics based on simulation results and discuss how sequence length interacts with base rate (unconditional probability of a student experiencing some state) and number of unique possible states a person might experience. Finally, we explore the effect of impossible loops on different transition metrics to determine the magnitude of spurious effects that are likely to be observed in these situations. Our software (written in Python) for transition metric calculations and simulations is publicly available for researchers to utilize and expand upon (Bosch & Paquette, 2020).

The analyses and contributions in this article are novel in several ways. This article is the first to discuss issues of sequence length and the effect of impossible loops across a variety of transition metrics. It is also the first to compare several widely-used transition metrics, which shows that many patterns are consistent across metrics. Comparison also reveals situations where apparently significant results may be due to unexpected properties of the metrics rather than meaningful patterns in data.

2. RELATED WORK

We consider several types of related work. First, we discuss work that exemplifies how transition metrics are commonly utilized to answer behavioral research questions. We then discuss work that examines the properties of transition metrics and research on sequences with impossible loops.

2.1. UTILIZING TRANSITION METRICS

Several different metrics have been utilized to study the transitions between states. Perhaps the simplest method is to measure the conditional probability of one state, given the previous state. This metric can be represented with the probabilities as edge weights in a graph of states (Figure 1), forming a Markov chain model (*MCM*). However, *MCM* probabilities are heavily influenced by the rate of occurrence of each state (the *base rate*), which can be an undesirable property for some research. For example, if state Y in Figure 1 occurs 90% of the time, the transition from X to Y is unsurprising, but if state Y occurs only 50% of the time, the $X \rightarrow Y$ transition is potentially indicative of a meaningful effect where X often leads to Y. This issue with *MCM* has motivated researchers to develop and utilize alternative transition metrics, including D'Mello's L (D'Mello & Graesser, 2012), lag sequential analysis (*LSA*; Faraone & Dorfman, 1987), and association measures such as Yule's Q (Walsh & Ollenburger, 2001).

Since their inception, transition metrics have been utilized primarily for behavioral research. In one of the earliest papers on transition metric methods, researchers explored transitions as a means to understand sequences of behavior across individuals (Bakeman & Dabbs, 1976). They discussed one example of undergraduate students having conversations, finding that students tended to look away before starting to speak to another person. In this application, behaviors of several individuals were combined into a single sequence of events; transition metrics then measured the conditional probability of one event following another, including events from different individuals.

Transition metrics are also effective for quantifying within-person phenomena. They have been used extensively to model the flow of one emotion to another, primarily during education-related activities such as learning computer programming (Bosch & D'Mello, 2017; Guia et al., 2013), medical training (Ocumpaugh et al., 2017), and problem solving (Baker et al., 2007), which has enabled the expansion of theories of emotion in learning (D'Mello & Graesser, 2012). Researchers have examined transitions and related phenomena for many different types of emotion data, such as continuous-valued estimates of valence or discrete states such as *bored* or *confused* (see Hamaker, Ceulemans, Grasman, & Tuerlinckx [2015] for a review). We focus here on metrics for transitions between discrete states in particular. In one such study, D'Mello & Graesser (2012) proposed a theoretical model of affect transitions during learning. The model suggested, for example, that confusion might be beneficial to learning if it was resolved (transitioning to engagement) or not beneficial if it was unresolved (transitioning to frustration), and was experimentally tested by applying transition metrics to sequences of emotions. They measured these transitions relative to chance with L, a transition metric which measures the rate of occurrence of transitions relative to their expected probability.

In related emotion research, Rodrigo (2011) measured transitions between emotional and cognitive states with the L metric in a game-based learning environment and found, for example, that boredom tended to persist while confusion was transitory and often led to engagement. These findings contribute to psychological theories of learning and also demonstrate how loops in a transition graph can be interpreted as "persistence," i.e., that a person's behavior is more likely to remain unchanged than to transition to another state. Similarly, Ocumpaugh et al. (2017) examined transitions between emotions while military trainees engaged with simulation software, retaining loops to study persistence.

Researchers have also examined transitions between other types of states, such as actions performed in software systems. For instance, Galyardt & Goldin (2015) examined metacognitive strategies represented by sequences of actions in a simulated intelligent tutoring system, while Bosch & D'Mello (2017) measured transitions between actions in a computer programming environment with emotions interspersed in the sequences of actions. Further examples of transition analyses include studying sequences of speech behaviors to predict alcohol drinking behavior (Gaume et al., 2008; Moyers et al., 2009), predicting antisocial behaviors (Dishion et al., 2004), and others in various education domains and beyond (Altermatt et al., 2002; Marion et al., 2003; Wuerker et al., 2001). These studies demonstrate some of the possibilities of transition metrics, which offer a perspective for examinations of behavior that is complementary to typical analyses of individual states.

2.2. Inspecting Transition Metric Methodologies

There are many possible transition metrics, including those designed specifically for the purpose and correlation measures that can be utilized as well. Here we briefly discuss some of the most relevant research on transition metrics and offer more details of the metrics themselves in the Method section.

One of the first metrics to be studied was *MCM* (Bakeman & Dabbs, 1976). Bakeman & Dabbs compared *MCM* to unconditional probabilities of two states occurring independently (rather than in sequence) as a means of providing context to the *MCM* values. Later, Sackett, Holm, Crowley, & Henkins (1979) provided a FORTRAN computer program to calculate the

MCM and unconditional probabilities, as well as a z-score that compared the two numbers, which is commonly referred to as lag sequential analysis (*LSA*). *LSA* was criticized in later publications for having somewhat unintuitive properties that might hinder interpretation (Bakeman et al., 1997). In particular, the magnitude of an *LSA* score is relative not only to the probability of the transition relative to random chance but also by the length of the sequence of data, which is often not an intended or expected effect.

Previous work has compared Yule's Q to various other transition metrics. Q is a monotonic transformation of the log odds ratio (see the Method section for details), which captures the odds of a transition occurring relative to random chance level. Bakeman, Mcarthur, & Quera (1996) noted Q (and thus monotonic functions of Q like log odds ratio) produced similar results in simulations of 100 transitions, where transition probabilities were systematically varied. Lloyd, Kennedy, & Yoder (2013) further analyzed Yule's Q with random simulations, finding that the correlation between Q and base rate (which dictates chance level) was $|r| < .09$. This corroborates the work of Bakeman et al. (1996) in establishing that Q is not biased by states that occur more or less frequently than others.

Our results also explore the effect of impossible loops, which has been explored for some metrics in previous work. D'Mello & Graesser (2012) noted that with frequent observations of emotional states, the probability of observing the same incident of one emotion multiple times is high, and thus loop probabilities are high. However, their objective was to measure the transitions between emotions, so they removed loops from the sequences before computing L, leaving only the transitions between states. As noted by researchers utilizing *LSA* and related approaches, however, it is necessary to modify calculations to account for situations where loops have been removed (Bakeman, 1983; Bakeman & Quera, 1995; Matayoshi & Karumbaiah, 2020). Modified calculations are needed because, after removing loops, the probability of transitions to all other states is increased a non-trivial amount. As noted in recent work focusing on L, this effect can influence the significance and even the direction of findings (Karumbaiah et al., 2019), solidifying the need for careful application of transition metric methods. Our work contributes to this literature by measuring the effect of loop removal on previously unexplored transition metrics and by exploring the relationship between loop removal and sequence length.

3. METHOD

The methods in this study consist of simulating sequences of states and evaluating these sequences with various transition metrics as well as applying the metrics to sequences collected in a computer-based learning environment. We examine the results (i.e., values of the metrics) to reveal key differences and similarities between metrics.

3.1. TRANSITION METRICS IN THIS STUDY

We describe metrics in terms of the cells of a 2×2 contingency table (Table 1) computed for a particular transition to be measured in a sequence (e.g., $X \rightarrow Y$), though the sequence may contain other states. Rows in the contingency table indicate counts of preceding states in all possible transitions, and columns indicate counts of states that follow. We refer to the counts in the cells of the table by A, B, C, and D in equations. In general, any measure of association between two variables in a contingency table can be used as a transition metric (e.g., Cohen's κ, correlation measures like φ); we focus on metrics that have been used in previous learning analytics research, including *MCM* (Dong & Biswas, 2017; Galyardt & Goldin, 2015; Jeong & Biswas, 2008), L (Baker et al., 2007; Bosch & D'Mello, 2017), L^* (Matayoshi & Karumbaiah, 2020), *LSA* (Chen et al., 2017; Yang et al., 2015), and *Q* (Molinari et al., 2013; Tompkins et al., 2013), and provide functionality for other association metrics in the accompanying software.

Table 1: Example 2x2 contingency table calculated from a hypothetical sequence of 201 observations of three or more states (*X*, *Y*, and others) for the transition from state *X* to state *Y*. Note that the sum of cells is 200 rather than 201 because it is unknown what follows the last state in the sequence, and thus no transition can be calculated.

n = 200		Next state (offset sequence)	
		State *Y*	**State is not *Y***
Current state (original sequence)	**State *X***	74 (*A*)	44 (*B*)
	State is not *X*	52 (*C*)	20 (*D*)

3.1.1. Markov-chain Model (MCM) Probability

Perhaps the most straightforward way to measure the propensity for transitions from state *X* to state *Y* is to calculate the conditional probability of *Y* given *X* (Equation 1), where the probability of a state occurring depends only on the one previous event (i.e., it satisfies the Markov property). *MCM* as a transition metric is straightforward to define and compute but slightly more difficult to interpret because it is influenced by differences in base rate. For example, an $X \rightarrow Y$ transition may appear to be occurring with unusual frequency simply because *Y* is especially common or uncommon.

$$MCM = P(Y|X) = \frac{A}{A+B} \tag{1}$$

3.1.2. D'Mello's *L*

L addresses the interpretability issues of *MCM* by subtracting the expected rate of occurrence for a particular transition so that values of 0 indicate a transition is occurring as often as would be expected in randomly-ordered data (Equation 2). The metric value is also scaled so that the maximum value is 1 regardless of state base rates. *L* has no lower bound, however, so negative values can have large magnitude and are thus less straightforward to interpret than positive values.

$$L = \frac{P(Y|X) - P(Y)}{1 - P(Y)} = \frac{\frac{A}{A+B} - \frac{A+C}{A+B+C+D}}{1 - \frac{A+C}{A+B+C+D}} \tag{2}$$

3.1.3. L*

The transition metrics we consider make the assumption that all transitions between states are possible. However, this is not always the case. Matayoshi & Karumbaiah (2020) explored the case where transitions are impossible because loops (transitions from one state to itself) have been removed – which might be done, for example, if loops were recorded in the original data due to a high sampling rate (e.g., when observing the same emotional event twice in a short period of time). L^* is a modified form of L that accounts for this specific case (Matayoshi & Karumbaiah, 2020). L^* follows the same equation as Equation 2, except that probabilities are computed from only the transitions $X \rightarrow Y$ where X differs from Y.

3.1.4. LSA

The interpretation of LSA is relatively straightforward because it produces a z score (Equation 7). Thus, for example, a result where $|LSA| > 1.96$ implies the transition being measured occurs significantly more or less often than expected by chance (two-tailed $p < .05$). However, the value of LSA is also influenced by sample size, since both effect size and significance are reflected in the value. Thus, the value of LSA is difficult to interpret as an effect size.

LSA, like the other transition metrics we explore here, is often calculated with a lag of 1, which means that transitions are measured between consecutive states. However, there is also some research in which larger lags are examined, allowing researchers to measure longer-term associations between actions. Hence, we also examine a larger lag value of 5 (referred to as *LSA-5*) as an additional example.

$$O_{XY} = observed\ X \rightarrow Y\ transitions = A \tag{3}$$

$$F_X = transitions\ from\ X = A + B \tag{4}$$

$$T_Y = transitions\ to\ Y = A + C \tag{5}$$

$$E_{XY} = expected\ X \rightarrow Y\ transitions = \frac{F_X T_Y}{A + B + C + D} \tag{6}$$

$$LSA = z = \frac{O_{XY} - E_{XY}}{\sqrt{E_{XY}\left(1 - \frac{F_X}{N}\right)\left(1 - \frac{T_Y}{N}\right)}} \tag{7}$$

3.1.5. Yule's Q

An odds ratio (OR) can be calculated as the odds of state Y following state X. Yule's Q (Yule, 1900) is a simple transformation of OR so that it ranges from -1 to 1, with 0 indicating random chance level (Equation 8) – unlike OR, which ranges from 0 to infinity with 1 indicating chance. Though its range and midpoint match those of common association measures like Cohen's κ, the values of Q can deviate considerably in certain cases. If $A = 0$, the value of D can change arbitrarily without influencing the result (and vice versa if $D = 0$). Similarly, if $B = 0$, the value of C has no influence on the result, and vice versa if $C = 0$. For example, if X transitions to Y in a sequence, but Y never transitions to another state (e.g., *XXYYY*), then $D = 0$ and the fact that X transitions to Y once ($A = 1$) makes no difference when measuring Q for the $X \rightarrow Y$ transition.

$$Q = \frac{OR - 1}{OR + 1} = \frac{AD - BC}{AD + BC} \tag{8}$$

3.2. SIMULATIONS

We generated simulated datasets to quantify the properties of transition metrics in precisely controlled conditions. In particular, we generated random datasets so that the expected result of calculating transition metrics for each dataset would be chance level (a null result). For each experiment, we generated 10,000 independent sequences. These datasets can be thought of as sequences of observations from 10,000 unique students in a study, for example, rather than 10,000 trials from one student (since each sequence is completely independent). We computed all transition metrics for each sequence after it was generated so that metrics would be calculated with exactly the same input data. Finally, we averaged each transition metric value over the 10,000 sequences to produce a final value for each metric.

3.2.1. Experiment 1 Simulations

Short sequences of states may produce misleading results for at least two reasons. First, transition metrics cannot be calculated in some short sequences, resulting in *invalid values*. This may occur, for example, when calculating Q: if $D = 0$ and $B = 0$, the denominator of Equation 8 is 0, and thus Q cannot be calculated. The probability of invalid values occurring by chance is higher for shorter sequences, since there are fewer opportunities to observe infrequent state transitions.

Second, shorter sequences may produce less accurate estimates of the strength of a transition and thus result in increased chances of finding outliers. For instance, a very short sequence of XYY may seem to indicate that X always transitions to Y (i.e., MCM for $X \rightarrow Y = 1.0$). However, the true probability may be much lower but not apparent due to the short length of the sequence. This issue is exacerbated in the presence of state imbalances, where one or more states infrequently occur even in long sequences. However, in experiment 1 simulations, we focus on the best-case scenario – when states are balanced – and note that issues arise even in this scenario.

Thus, it is important to determine how long a sequence of observed states must be to successfully measure the desired state transitions and avoid invalid values. We conducted simulations varying sequence length to quantify the relationships between sequence length and the number of invalid values encountered and between sequence length and the maximum transition metric value observed.

We calculated the maximum metric value calculated from among the means across 10,000 sequences. For example, with only two unique states, there are two possible transitions ($X \rightarrow Y$ and $Y \rightarrow X$); we calculated the mean $X \rightarrow Y$ transition across all sequences, and likewise for $Y \rightarrow X$, then found the maximum of those two means. Thus, these maximum values did not simply represent outliers from among the 10,000 sequences; rather, they are stable across many sequences.

These simulations required specifying the number of unique states from which to sample to create sequences. In practice, the number of unique states varies across domains. For example, researchers in one study examined transitions between seven different emotions (Baker et al., 2007); another study considered transitions between ten different self-regulated learning behaviors (Witherspoon et al., 2008), and in another study, researchers considered transitions between six emotions and eight behaviors (14 total states; Bosch & D'Mello, 2017). A dataset with a large number of states (versus a small number) will have more possible transitions between states and a lower base rate of occurrence for some or all states. Consequently, we expect that longer sequences would be needed with more states to achieve low error, given that there is less available evidence for each transition in a given sequence length. We contrasted both the simplest possible case (two states) and a slightly more complex case (four states) in experiment 1 simulations. These examples serve to illustrate the expected trend and the magnitude of potential errors due to short sequences, even with few states. We then also briefly describe how this trend continues with seven states, chosen as a realistic number to match one previous study (Ocumpaugh et al., 2017). These choices of sequence lengths and number of states do not fully cover the range of possibilities that could be encountered in research; however, they do illustrate trends that can be used to extrapolate expectations for research with the same order of magnitude of states and sequence lengths.

3.2.2. Experiment 2 Simulations

We ran a second set of simulations to explore the effect of impossible loops on transition metric values. In these simulations, we considered three possible states because it is the simplest non-trivial case. The case of two possible states is trivial because, without loops, each state is left with only the possibility to transition to the other state and vice versa. For example, *XYYXXXY* becomes *XYXY* without loops, and $X \rightarrow Y$ transition probability will always equal 1.0 (as will $Y \rightarrow X$).

State imbalance (differences in base rates of occurrence for different states) is also important to consider for cases where loops are removed, since removal can influence base rates. Thus, for these simulations, we considered sequences of three states with 50%, 25%, and 25% base rates of occurrence. We measured the values of transition metrics along with their standard deviations (across all 10,000 random sequences) to determine how metrics differed with and without loops. These sequences were each 100 states long, which we chose to avoid issues that can arise with short sequences (see experiment 1 results).

3.3. Computer-based Learning Environment Data (Experiment 3)

Randomly-generated sequences provide the opportunity to examine transition metrics in a situation where the null hypothesis is known to be true. That is, transition metrics should produce null results. Conversely, data collected in real-world learning situations are generated by students (or teachers) who are, presumably, not performing actions or having experiences at random. We examine one real-world dataset from a computer-based learning environment to determine how experimental results may differ with such data. We expect transition metric values to converge as sequence length increases, but not necessarily toward a null value.

The data we examined came from the Educational Process Mining (EPM) dataset (Vahdat et al., 2014). EPM data were collected from 99 students in a digital electronics course at the University of Genoa. Students learned in a computer-based learning environment called *Deeds* (Digital electronics education and design suite), which allowed students to design and simulate circuits, read learning materials, take notes, and do other learning-related activities. Deeds recorded (i) the sequence of exercise IDs students worked on, (ii) the sequence of learning activities within each exercise, and (iii) the number of actions within each activity, such as the number of mouse clicks and keystrokes. We examined sequences (i) and (ii) from these data, which serve as examples of key transition metric cases. The exercise sequence data included many loops (95.8% of transitions were loops) because each activity within an exercise was recorded as a separate step, and exercise IDs were repeated for each step. Hence, we removed loops from this sequence since loops dominate the sequence yet provide no insight into the order of exercises students did. The activity sequence data included relatively few loops, on the other hand (14.3% of transitions), and loops are meaningful transitions for the activity sequence because they indicate that a student transitioned from one particular activity type in an exercise to the same activity type in a different exercise.

We examined the effect of sequence length on transition metric values for these sequences by extracting subsequences from the beginning of each sequence, varying in length from 5 to 50 steps. We also calculated transition metrics on the full sequences for each student, which ranged in length from 87 to 1087 actions ($M = 442.9$, $SD = 162.9$). These full-length sequences make it possible to determine whether trends observed in the initial 50 sequence steps converge for long sequences.

4. RESULTS

We present results to answer two general research questions: 1) how does sequence length impact transition metric values? and 2) how do impossible loops impact different transition metric values?

4.1. EXPERIMENT 1A: INVALID METRIC VALUES VS. SEQUENCE LENGTH

We first examined the occurrence of invalid results (i.e., “not a number” results) that were produced with random state sequences of varying lengths. Not all metrics produce invalid results for the same sequences. For example, the value of Q will be invalid if the denominator is zero in Equation 8, which occurs when either A or D is zero and either B or C is zero. Conversely, simple *MCM* probability will be invalid only when A and B are both zero.

Figure 2 shows the trend in the proportion of invalid values for each metric versus sequence length. In general, shorter sequences were more susceptible to producing invalid values for these randomly-generated sequences. Additionally, invalid values were more common with a higher number of states (four, rather than two). *LSA* with lag 5 (i.e., *LSA-5*) followed the same pattern as *LSA* (lag 1, by default), shifted by 4 units on the x-axis. This was expected, given that sequences were random and were functionally shorter by 4 states when lag increased from 1 to 5. As results show, shorter sequences resulted in more invalid values.

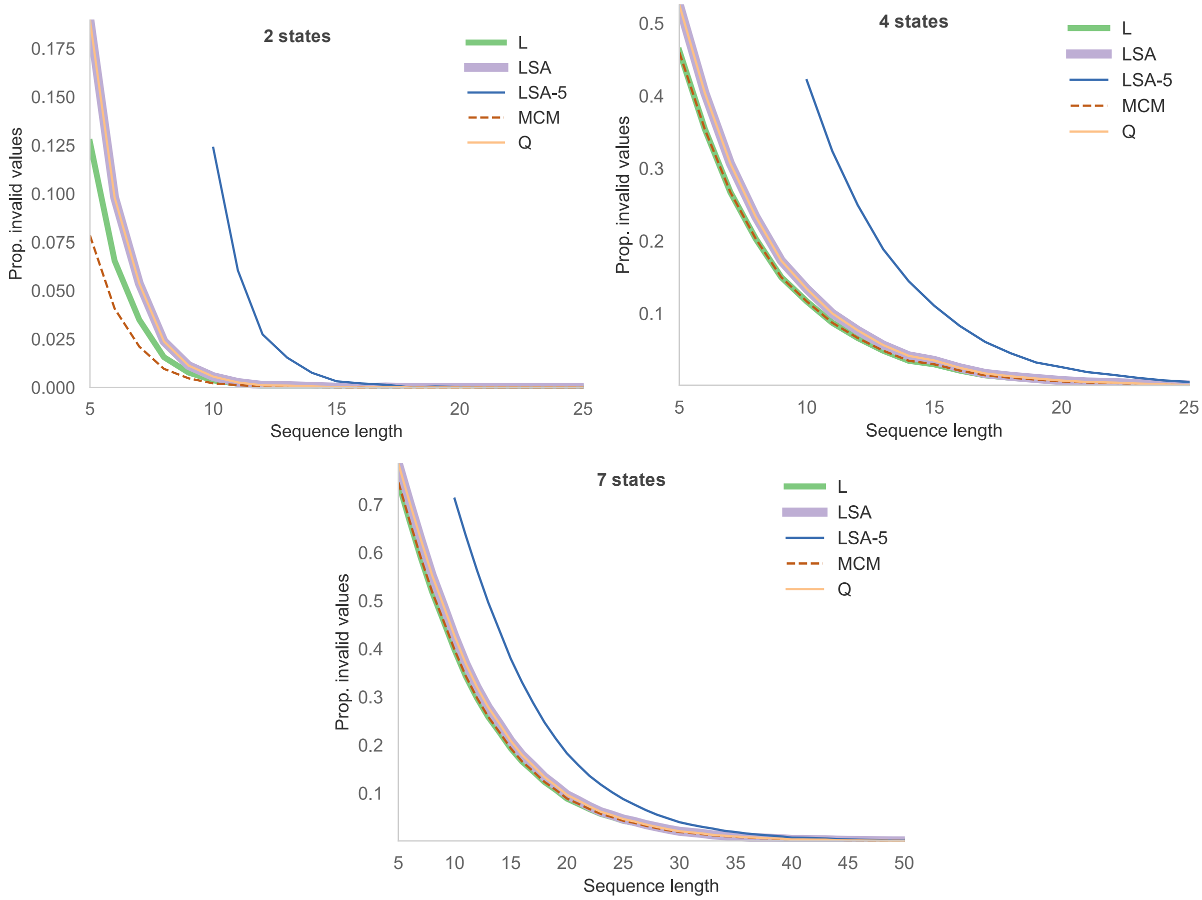

Figure 2: Proportion of invalid values for transition metrics as a function of sequence length for random sequences of two states (top left), four states (top right), and seven states (bottom). Overlapping lines are shown with varying styles to avoid occultation. Note differences in *x*- and *y*-axes given that more states resulted in substantially more invalid values.

4.2. Experiment 1b: Maximum Metric Value vs. Sequence Length

Of the transition metrics we consider in this paper, all have a chance level of 0 except *MCM*. That is, given random sequences, they should (on average) not result in values notably above or below 0. For example, $Q = 0$ indicates no association between two variables. *MCM* is different since it is a probability; the chance-level value of $P(Y \mid X)$ is the probability of Y. If there are two equally probable states (X and Y) in a sequence, the expected value of *MCM* for $X \rightarrow Y$ transitions in random sequences is 1 / 2 = .5. With four equally probable states, the expected value of *MCM* for $X \rightarrow Y$ transitions is 1 / 4 = .25, and with seven it is 1 / 7 = .143.

We computed the maximum value observed among all transitions in a simulation to determine whether there were systematic (averaged across all iterations) deviations from 0. Figure 3 shows these maximum values relative to sequence length. Results indicated that there were indeed cases where short sequences can produce seemingly significant results. This effect was more notable with fewer states in most cases except perhaps Q, which had more negative values for larger numbers of states. This occurs because Q subtracts transitions between state pairs that are not the two states of interest for a particular transition (*BC* in Equation 8). With more states, there are many such transitions, and Q is thus often negative, as seen in Figure 3 right and bottom.

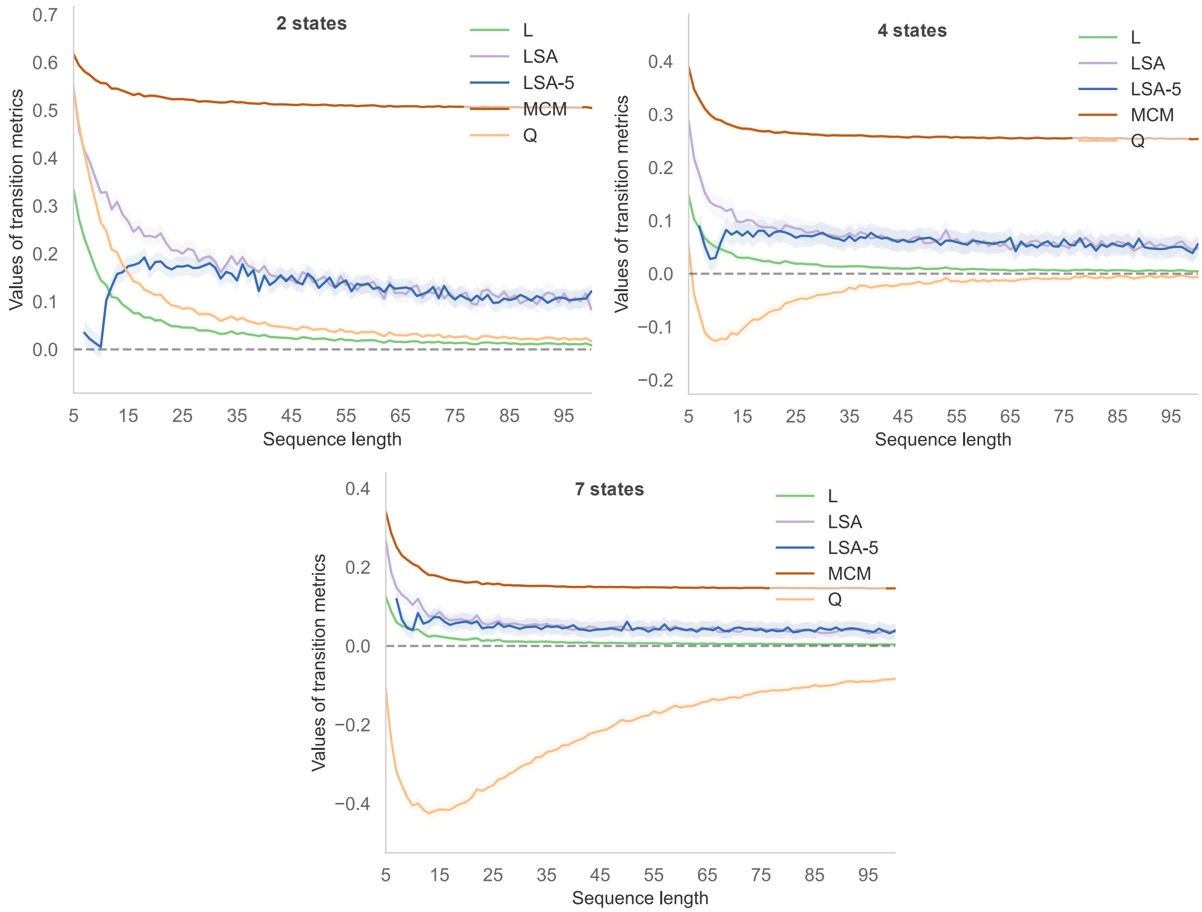

Figure 3: Maximum average transition value observed over all simulations for varying sequence lengths and two (top left) versus four (top right) versus seven (bottom) states, with 95% confidence intervals indicated by shading. All metrics converge toward 0 except MCM, which converges toward 0.5 for two states, 0.25 for four states, and 0.143 for seven states. LSA converges extremely slowly because longer sequences increase the magnitude of LSA results, including spurious results.

4.3. Experiment 2: Metric Values in Sequences without Loops

Sequences where loops are impossible (either because of the nature of the data collection environment or because of data preprocessing steps) have different probabilities of transition from one state to another than would otherwise be expected. We investigated both of these aspects with additional simulations where we calculated transition metrics with and without loops (by removing loops – i.e., repeated states in the sequence – wherever they occurred) on the same set of randomly-generated state sequences of length 100. We did not consider lags other than 1 for *LSA*, given that experiment 1 results showed that larger lags essentially shortened the sequences but did not otherwise differ from lag 1 results.

Figure 4 shows the effect of having sequences where loops are not possible. The initial base rate of each state was directly tied to probability of transition to that state (e.g., probability of transition from any state to X was approximately .250 since the base rate of X was 25%). With no loops (Figure 4 right), the base rates of states themselves were influenced because fewer loops occurred by random chance for less common states, and thus fewer instances of the uncommon states were removed compared to the more common state.

It is evident from Figure 4 that metrics change without loops and that transition probabilities no longer match base rates of the destination states. Furthermore, transition probabilities do not match the new base rates of destination states even after ignoring the initial state (e.g., ignoring X in an $X \rightarrow Z$ transition and calculating the Z base rate as Z / [Z + Y] = 40% / 70% = .571). For other transition metrics with clearly defined chance levels, related phenomena occur; some transitions were apparently well above chance-level (0, for L, LSA, and Q), despite the fact that sequences were generated randomly. Each metric consists of calculating some expectation for how often transitions should occur, and when loops are impossible, the expected value is no longer correct (it is 0 for the loop and thus slightly higher for all other transitions). L* was unaffected, however, which is as expected since it corrects for this specific problem (Matayoshi & Karumbaiah, 2020).

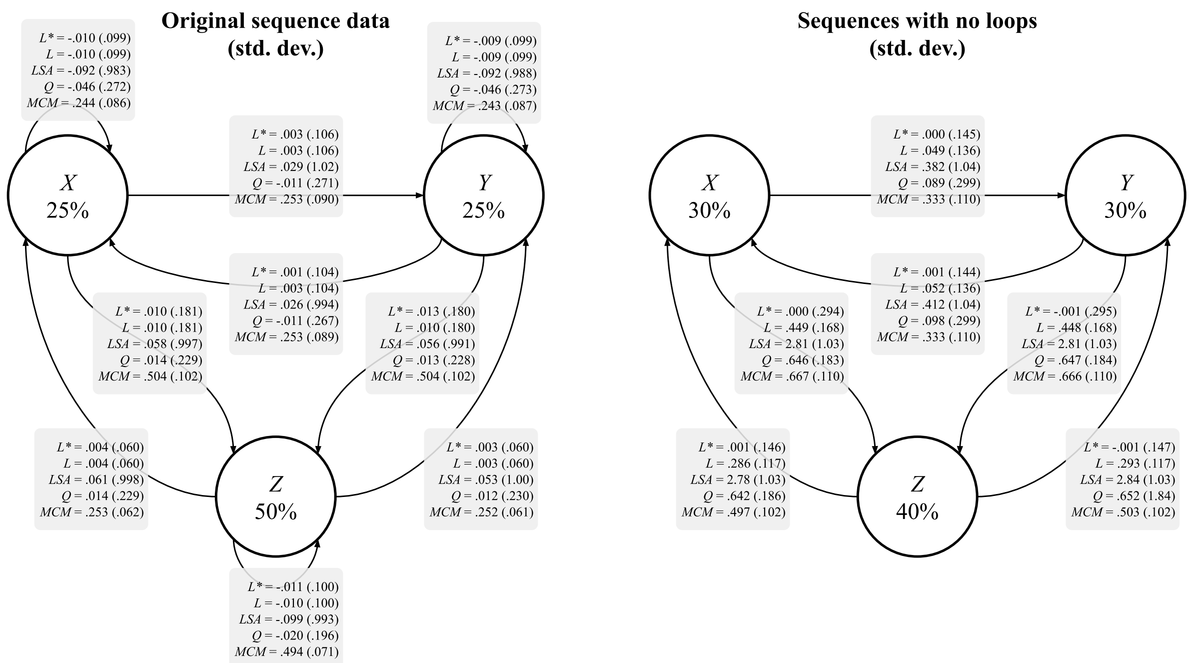

Figure 4: Transition metric values before (left) and after (right) removing loops (transitions from a state to the same state) in random sequences of length 100, with three possible states (X, Y, and Z). Percentages inside circles indicate base rates of each state. Numbers in parentheses are standard deviations for each metric.

4.4. EXPERIMENT 3: ACTIVITY SEQUENCES IN COMPUTER-BASED LEARNING

In experiment 3, we explored the effects of sequence length on transition metric values with data collected in the wild from students using a computer-based learning environment. This experiment shows the importance of sequence length outside of simple simulations. However, because of the non-random nature of activity sequences in real learning contexts, we do not expect transition metrics to converge to a specific null value (i.e., the mean base rate for MCM or 0 for all others). We might expect mean values across all possible transitions to approach the null value for a metric, but this behavior is not guaranteed; for example, the range of L is $(-\infty, 1]$ and thus can have negative outliers that draw the mean far from zero. We also examined the transitions with the minimum value and the maximum value, in addition to the mean across transitions, to provide additional perspectives on how metrics might behave differently in these data. Note that these minimums and maximums were the lowest and highest mean values of a transition obtained after averaging each transition over all students. Hence, they represent transitions that were consistently low or high across students: for example, the maximum-value transition is the maximum mean across students, rather than the mean of the maximum within each student, which might be a different transition for each student.

Trends in the results for sequences of activities students performed (e.g., use a text editor, use an electronics simulator) indicated that the minimum-value transition tended to converge toward the lowest possible value as sequence length increased, the maximum-value transition trended toward the highest possible value, and some means converged toward approximately a null value. For example, the minimum-value transition converged toward 0 for *MCM* and toward -1 for *Q* as sequence length increased (Figure 5 top left), which are the minimum possible values for those metrics. This indicates that there was at least one transition that almost never happened, and this was evident from sequences even shorter than 10 for those metrics. Conversely, *LSA* – which has no minimum possible and is a *z*-score that depends on sequence length – trended continually lower, while *L* (which also has no minimum possible but is not relative to sequence length) converged toward a negative value. These trends extend in the full-length values in Figure 5 (lower right), where it is apparent that, on average, the minimum and maximum *LSA* values continued to diverge while minimum *MCM* reached its lowest possible value (0), as did *Q* (-1), and maximum *L*, *MCM*, and *Q* were all near (though not quite at) the maximum possible value of 1. These trends were also apparent in very short sequences, like the minimum-value transition. Conversely, mean metric values for all transition metrics were close to the chance-level value of 0 for *L* and *LSA*. Similarly, mean *MCM* was close to its chance level of .125 (because students engaged in 8.02 unique activity types, on average). These results indicate that, on average, the transitions above and below chance level tended to cancel out, though that trend required relatively long sequences to consistently observe, especially for *MCM*.

LSA with a larger lag (i.e., *LSA-5*) results differed in notable ways for the EPM dataset, unlike for the randomly-generated sequences. Like *LSA* (lag 1), the maximum- and minimum-value transitions for *LSA-5* (Figure 5) trended increasingly positive and negative, respectively, as sequence length increased. However, *LSA-5* maximum and minimum values were both smaller than *LSA*, including in the full-length sequences. This suggests that there were clear above- and below-chance transitions for lags 1 and 5 but that the trends were clearer for lag 1. Intuitively, this pattern might be expected: the activity directly preceding the current activity is likely more related to the current activity than the fifth most recent activity is.

Results from the sequence of exercises that students worked on, with loops removed, showed primarily that students tended not to revisit previous exercises, and worked in a predictable order (Figure 6). Maximum and minimum transitions converged quickly toward the highest and lowest possible values for metrics with clearly defined limits, indicating that there were some transitions that almost always happened and some that almost never did. However, the mean metric values (Figure 6 lower left) show that the mean was slow to converge as sequence length increased, and thus there were some transitions that happened infrequently – but more often than never. More notably, the jagged nature of the mean transition metric values suggests that even longer sequences up to 50 states were insufficient to provide a stable estimate of transitions with loops removed. This is unsurprising given that many of the transitions consisted of loops (see section 3.3). Even the full-length sequences in Figure 6 (lower right) may not have all converged, as evidenced by the fact that the magnitude of minimum *LSA* was smaller than 1.96, which corresponds to the typical $p = .05$ threshold given that *LSA* is a *z*-score. Consequently, this result illustrates the importance of considering the effect of loop removal on sequence length, since even long sequences may become too short to produce reliable results when loops are removed.

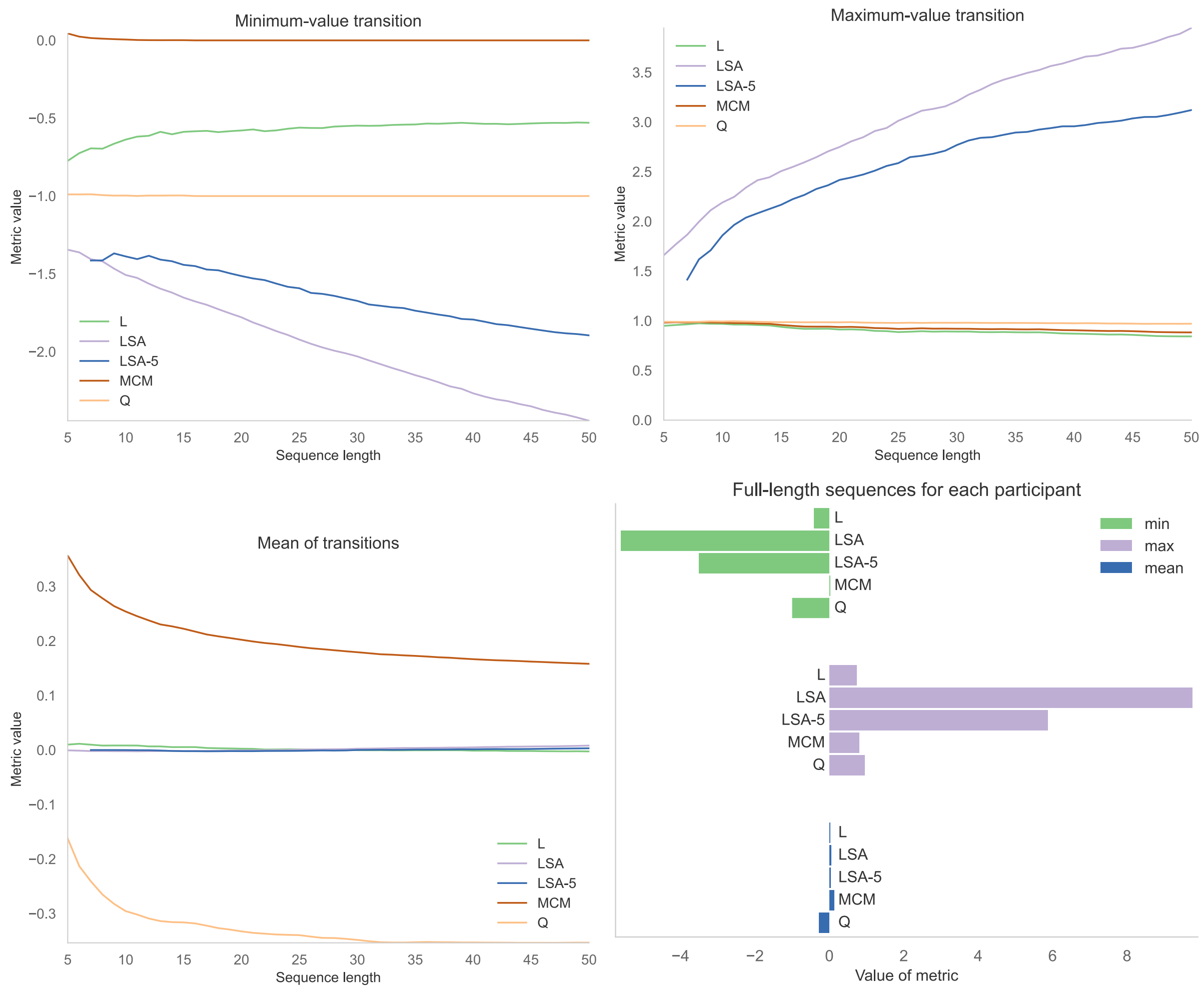

Figure 5: Minimum (top left), maximum (top left), and mean (bottom left) transition metric values observed across all possible **transitions between activities** recorded in the EPM dataset. Full-length sequences (lower right) show how observed sequence length trends extend to sequences hundreds of steps long.

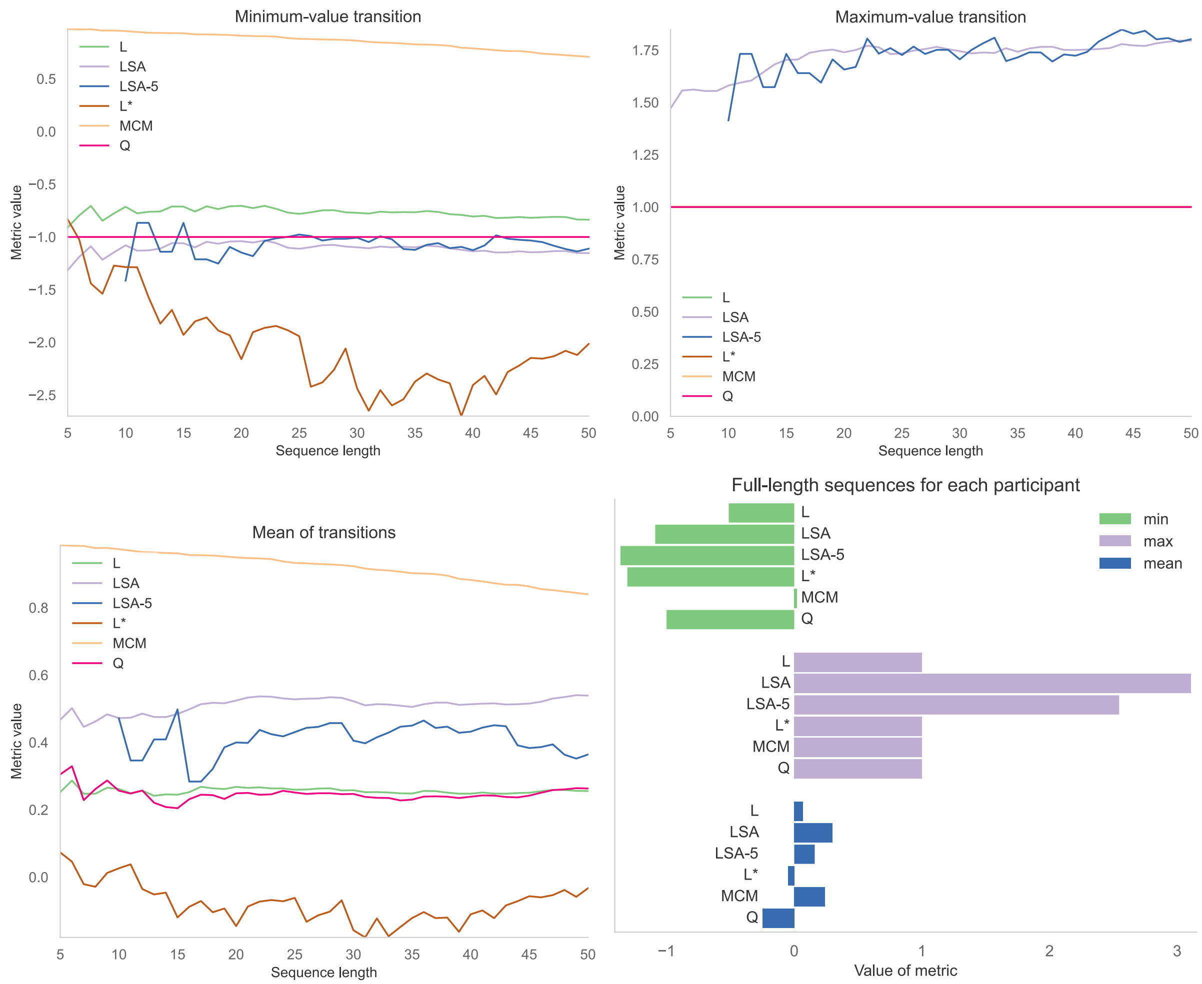

Figure 6: Minimum (top left), maximum (top right), and mean (bottom left) transition metric values for the **sequence of exercises** students worked on in the EPM dataset, with loops in the sequence removed. L, L^*, and MCM all have maximum values of 1 and are thus occluded by Q in the top right subfigure. Full-length sequences (lower right) illustrate trends in the best-case scenario where sequences are as long as possible, i.e., they include the entire dataset.

5. Discussion

We were interested in the properties of state transition metrics, especially as they relate to sequence length. To study transition metrics, we first generated random state sequences to simulate various scenarios that occur in research, including short sequences and sequences with impossible transitions. We found several situations where results appeared to be above chance levels, despite the fact that sequences were randomly generated. Analyses of real-world data also showed the importance of having long input sequences. These results offer some insight into scenarios that should be avoided when applying these metrics to study sequential data so that spurious conclusions can be avoided.

5.1. MINIMUM SEQUENCE LENGTH

Sequence length impacted the number of invalid transition metric values observed in our results. These invalid values are undesirable for two reasons. First, they reduce the number of estimates of transition metric values that are available for analyses (i.e., reduced statistical power when testing whether transitions occur significantly more or less frequently than chance). Second, the presence of a notable proportion of invalid values for a specific sequence length suggests that the other sequences of the same length – which did yield valid values – were based on low-power estimates of the parameters in the equations (e.g., the expected probability of a particular state occurring).

We expected invalid values would be more common for shorter sequences, sequences with more possible states, and sequences with imbalanced base rates, because each of these scenarios increases the chances of encountering sequences with few (or even zero) transitions involving some states. Indeed, we found that shorter sequences yielded more invalid values in the random sequences we sampled, especially with more possible states. We suggest that researchers should consider the possibility of invalid values when deciding how long state sequences need to be during experiment design and analyses. Our simulation code is publicly available (Bosch & Paquette, 2020), which allows easy manipulation of the number of states and plotting invalid values across a range of sequence lengths. Behavioral researchers can thus generate sequences for the expected number of states in their work and observe the proportion of invalid values; for example, from the results of this paper, we might suggest sequence length should be at least 10 for 2 possible states, at least 20 for 4 possible states, and at least 35 for 7 possible states (Figure 2). In general, there are $n \times (n + 1)$ transitions between n states, including loops. Hence, we would expect the minimum necessary sequence length to grow quadratically with the number of states, so that each transition has some evidence from which to calculate its propensity. Indeed, this appears to be the pattern observed in the results. However, uncommon states would also reduce the available evidence for transitions to or from that state, which should be taken into account as well.

We also found that shorter sequences were more likely to result in spurious above-chance transition estimates. This issue was more pronounced with fewer states (two versus four) but was apparent for both cases (Figure 3). Overall, these results suggest that long sequences are needed to avoid spurious findings. Depending on metric and number of states, sequence length may need to be in excess of 50 to avoid spurious results over 0.1 (a “small effect” for some measures; Cohen, 1988). Our experiments with real-world data (Figure 6) highlight the fact that sequence length after loop removal may be substantially shorter, and thus even long sequences can produce noisy results once shortened due to loop removal.

Sequences may also vary in length in most practical applications. For example, students might interact with a computer-based learning environment for differing amounts of time or perform different numbers of actions within such environments. Since transition metrics are calculated per student to avoid issues with statistical dependencies, it is important to examine sequence length for each student to determine whether each meets a reasonable minimum. Alternatively, the *LSA* metric might be preferable in situations where some students have short sequences, because *LSA* penalizes shorter sequences by its nature.

5.2. IMPOSSIBLE LOOPS AND LOOP REMOVAL

We expected that situations where loops (i.e., self-transitions, state persistence) are impossible would result in increased transition metric values for transitions to other states, since the probability of transition to those other states would be higher. This was indeed the case for all metrics. Accounting for this effect is complicated by the influence of base rates; infrequent states are less affected by the fact that loops within those states are impossible. For situations where impossible loops are caused by loop removal, it is possible to correct for these base rate effects, at least for one metric (i.e., L^* correcting for L). Our results show that all transition metrics we explored were affected by this issue, apart from the L^* metric designed specifically to solve this issue. Moreover, results from the real-world computerized learning environment we investigated showed that results may be unstable even with seemingly long sequences of behaviors if loop removal results in dramatically shortened sequences. Hence, it is also crucial to consider the proportion of states in a sequence that are likely to be removed when doing loop removal.

However, there are also situations in which transitions other than loops may be impossible. For example, students might use a computerized learning environment in which it is only possible to view a hint after attempting to solve a problem, but not directly after starting the problem. In this case, the *show problem → hint* transition is impossible, and thus the expected probability of transitions from other states to *hint* should be increased to maintain the sum of probabilities being 1. This is a broader case of the impossible loop situation, which merits further research in the future. For example, an alternative approach to transition analysis might be to treat each transition as a probability, like *MCM*, but to estimate these via Bayesian methods that could account for any impossible transitions (whether loops or not) by specifying prior distributions for each transition.

5.3. CONCLUDING REMARKS

We were motivated to examine the properties of state transition metrics, given their importance in analysis of student behaviors and emotions. To do so, we conducted simulations with random sequences of states, uncovering situations where such metrics produce misleading results (in short sequences or with impossible loops) and suggested ways to avoid or correct for these situations. Although our findings are intuitive (i.e., short sequences cause problems), the results still have potential to influence behavioral research by providing specific guidance regarding the length of sequences needed, and our open-source transition metric calculation and simulation software will make application of metrics more straightforward. Our software supports simulation for prospective analyses before data collection begins, calculation of transition metrics from collected data, and integration as a library into other software packages. Ultimately, we hope that these methods will be applied to develop a more accurate understanding of student behavior.

REFERENCES

ALTERMATT, E. R., POMERANTZ, E. M., RUBLE, D. N., FREY, K. S., & GREULICH, F. K. (2002). Predicting changes in children's self-perceptions of academic competence: A naturalistic examination of evaluative discourse among classmates. *Developmental Psychology*, *38*(6), 903–917. https://doi.org/10.1037/0012-1649.38.6.903

ANDRES, A., OCUMPAUGH, J., BAKER, R. S., SLATER, S., PAQUETTE, L., JIANG, Y., BOSCH, N., MUNSHI, A., MOORE, A. L., & BISWAS, G. (2019). Affect sequences and learning in Betty's Brain. In C. Brooks, R. Ferguson, & H. U. Hoppe (Eds.), *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19)* (pp. 383–390). ACM. https://doi.org/10.1145/3303772.3303807

BAKEMAN, R. (1983). Computing lag sequential statistics: The ELAG program. *Behavior Research Methods & Instrumentation*, *15*(5), 530–535. https://doi.org/10.3758/BF03203700

BAKEMAN, R., & DABBS, J. M. (1976). Social interaction observed: Some approaches to the analysis of behavior streams. *Personality and Social Psychology Bulletin*, *2*(4), 335–345. https://doi.org/10.1177/014616727600200403

BAKEMAN, R., MCARTHUR, D., & QUERA, V. (1996). Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared. *Behavior Research Methods, Instruments, & Computers*, *28*(3), 446–457. https://doi.org/10.3758/BF03200524

BAKEMAN, R., MCARTHUR, D., QUERA, V., & ROBINSON, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*(4), 357–370.

BAKEMAN, R., & QUERA, V. (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, *118*(2), 272–284. https://doi.org/10.1037/0033-2909.118.2.272

BAKER, R. S., D'MELLO, S. K., RODRIGO, MA. M. T., & GRAESSER, A. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

BAKER, R. S., RODRIGO, MA. M. T., & XOLOCOTZIN, U. E. (2007). The dynamics of affective transitions in simulation problem-solving environments. In A. C. R. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction* (pp. 666–677). Berlin, Heidelberg: Springer.

BISWAS, G., SEGEDY, J. R., & BUNCHONGCHIT, K. (2016). From design to implementation to practice—A learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education*, *26*(1), 350–364. https://doi.org/10.1007/s40593-015-0057-9

BOSCH, N., & D'MELLO, S. K. (2017). The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education*, *27*(1), 181–206. https://doi.org/10.1007/s40593-015-0069-5

BOSCH, N., & PAQUETTE, L. (2020). *Transition metrics simulation code* (1.0.0) [Computer software]. https://doi.org/10.5281/zenodo.3711563

CHEN, B., KNIGHT, S., & WISE, A. F. (2018). Critical issues in designing and implementing temporal analytics. *Journal of Learning Analytics*, *5*(1), 1–9. https://doi.org/10.18608/jla.2018.53.1

CHEN, B., RESENDES, M., CHAI, C. S., & HONG, H.-Y. (2017). Two tales of time: Uncovering the significance of sequential patterns among contribution types in knowledge-building discourse. *Interactive Learning Environments*, *25*(2), 162–175. https://doi.org/10.1080/10494820.2016.1276081

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.

DISHION, T. J., NELSON, S. E., WINTER, C. E., & BULLOCK, B. M. (2004). Adolescent friendship as a dynamic system: Entropy and deviance in the etiology and course of male antisocial behavior. *Journal of Abnormal Child Psychology*, *32*(6), 651–663. https://doi.org/10.1023/B:JACP.0000047213.31812.21

D'MELLO, S. K., & GRAESSER, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

DONG, Y., & BISWAS, G. (2017). An extended learner modeling method to assess students' learning behaviors. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, 302–305.

FARAONE, S. V., & DORFMAN, D. D. (1987). Lag sequential analysis: Robust statistical methods. *Psychological Bulletin*, *101*(2), 312–323. https://doi.org/10.1037/0033-2909.101.2.312

GALYARDT, A., & GOLDIN, I. (2015). Evaluating simplicial mixtures of Markov chains for modeling student metacognitive strategies. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative Psychology Research* (pp. 377–393). Springer. https://doi.org/10.1007/978-3-319-07503-7_24

GAUME, J., GMEL, G., FAOUZI, M., & DAEPPEN, J.-B. (2008). Counsellor behaviours and patient language during brief motivational interventions: A sequential analysis of speech. *Addiction*, *103*(11), 1793–1800. https://doi.org/10.1111/j.1360-0443.2008.02337.x

GUIA, T. F. G., RODRIGO, MA. M. T., DAGAMI, M. M. C., SUGAY, J. O., MACAM, F. J. P., & MITROVIC, A. (2013). An exploratory study of factors indicative of affective states of students using SQL-Tutor. *Research and Practice in Technology Enhanced Learning*, *8*(3), 411–430.

HAMAKER, E. L., CEULEMANS, E., GRASMAN, R. P. P. P., & TUERLINCKX, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

JEONG, H., & BISWAS, G. (2008). Mining student behavior models in learning-by-teaching environments. In R. S. Baker, T. Barnes, & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 127–136). International Educational Data Mining Society.

KARUMBAIAH, S., BAKER, R. S., & OCUMPAUGH, J. (2019). The case of self-transitions in affective dynamics. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 172–181). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_15

KNIGHT, S., WISE, A. F., & CHEN, B. (2017). Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics*, *4*(3), 7–17. https://doi.org/10.18608/jla.2017.43.2

LARSON, R., & CSIKSZENTMIHALYI, M. (1983). The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, *15*, 41–56.

LLOYD, B. P., KENNEDY, C. H., & YODER, P. J. (2013). Quantifying contingent relations from direct observation data: Transitional probability comparisons versus Yule's Q. *Journal of Applied Behavior Analysis*, *46*(2), 479–497. https://doi.org/10.1002/jaba.45

Marion, S. D., Touchette, P. E., & Sandman, C. A. (2003). Sequential analysis reveals a unique structure for self-injurious behavior. *American Journal on Mental Retardation*, *108*(5), 301–313. https://doi.org/10.1352/0895-8017(2003)108<301:SARAUS>2.0.CO;2

Matayoshi, J., & Karumbaiah, S. (2020). Adjusting the L statistic when self-transitions are excluded in affect dynamics. *Journal of Educational Data Mining*, *12*(4), 1–23. https://doi.org/10.5281/zenodo.4399681

McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective transitions in narrative-centered learning environments. *Educational Technology & Society*, *13*(1), 40–53.

McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2008). Affective transitions in narrative-centered learning environments. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS 2008)* (pp. 490–499). Springer. https://doi.org/10.1007/978-3-540-69132-7_52

Molinari, L., Mameli, C., & Gnisci, A. (2013). A sequential analysis of classroom discourse in Italian primary schools: The many faces of the IRF pattern. *British Journal of Educational Psychology*, *83*(3), 414–430. https://doi.org/10.1111/j.2044-8279.2012.02071.x

Moyers, T. B., Martin, T., Houck, J. M., Christopher, P. J., & Tonigan, J. S. (2009). From in-session behaviors to drinking outcomes: A causal chain for motivational interviewing. *Journal of Consulting and Clinical Psychology*, *77*(6), 1113–1124. https://doi.org/10.1037/a0017189

Ocumpaugh, J., Andres, J. M., Baker, R. S., DeFalco, J., Paquette, L., Rowe, J., Mott, B., Lester, J., Georgoulas, V., Brawner, K., & Sottilare, R. (2017). Affect dynamics in military trainees using vMedic: From engaged concentration to boredom to confusion. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 238–249). Springer. https://doi.org/10.1007/978-3-319-61425-0_20

Ocumpaugh, J., Baker, R. S., & Rodrigo, Ma. M. T. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 technical and training manual*. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Rodrigo, Ma. M. T. (2011). Dynamics of student cognitive-affective transitions during a mathematics game. *Simulation & Gaming*, *42*(1), 85–99. https://doi.org/10.1177/1046878110361513

Sackett, G. P., Holm, R., Crowley, C., & Henkins, A. (1979). A FORTRAN program for lag sequential analysis of contingency and cyclicity in behavioral interaction data. *Behavior Research Methods & Instrumentation*, *11*(3), 366–378. https://doi.org/10.3758/BF03205679

Tompkins, V., Zucker, T. A., Justice, L. M., & Binici, S. (2013). Inferential talk during teacher–child interactions in small-group play. *Early Childhood Research Quarterly*, *28*(2), 424–436. https://doi.org/10.1016/j.ecresq.2012.11.001

Vahdat, M., Oneto, L., Ghio, A., Donzellini, G., Anguita, D., Funk, M., & Rauterberg, M. (2014). A learning analytics methodology to profile students behavior and explore interactions with a digital electronics simulator. In C. Rensing, S. de Freitas, T. Ley, & P. J. Muñoz-Merino (Eds.), *Proceedings of the 9th European Conference on Technology*

Enhanced Learning (EC-TEL) (pp. 596–597). Springer International Publishing. https://doi.org/10.1007/978-3-319-11200-8_87

WALSH, A., & OLLENBURGER, J. (2001). *Essential statistics for the social and behavioral sciences: A conceptual approach*. Prentice Hall.

WITHERSPOON, A. M., AZEVEDO, R., & D'MELLO, S. K. (2008). The dynamics of self-regulatory processes within self-and externally regulated learning episodes during complex science learning with hypermedia. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS 2008)* (pp. 260–269). Springer. https://doi.org/10.1007/978-3-540-69132-7_30

WUERKER, A. K., HAAS, G. L., & BELLACK, A. S. (2001). Interpersonal control and expressed emotion in families of persons with schizophrenia: Change over time. *Schizophrenia Bulletin*, *27*(4), 671–686. https://doi.org/10.1093/oxfordjournals.schbul.a006906

YANG, T.-C., CHEN, S. Y., & HWANG, G.-J. (2015). The influences of a two-tier test strategy on student learning: A lag sequential analysis approach. *Computers & Education*, *82*, 366–377. https://doi.org/10.1016/j.compedu.2014.11.021

YULE, G. U. (1900). On the association of attributes in statistics: With illustrations from the material of the childhood society. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *194*(252–261), 257–319. https://doi.org/10.1098/rsta.1900.0019