Using Marginal Models to Adjust for Statistical Bias in the Analysis of State Transitions

Jeffrey Matayoshi
McGraw Hill ALEKS
Irvine, California, USA
jeffrey.matayoshi@aleks.com

Shamya Karumbaiah
University of Pennsylvania
Philadelphia, Pennsylvania, USA
shamya@upenn.edu

ABSTRACT

Many areas of educational research require the analysis of data that have an inherent sequential or temporal ordering. In certain cases, researchers are specifically interested in the transitions between different states—or events—in these sequences, with the goal being to understand the significance of these transitions; one notable example is the study of affect dynamics, which aims to identify important transitions between affective states. Unfortunately, a recent study has revealed a statistical bias with several metrics used to measure and compare these transitions, possibly causing these metrics to return unexpected and inflated values. This issue then causes extra difficulties when interpreting the results of these transition metrics. Building on this previous work, in this study we look in more detail at the specific mechanisms that are responsible for the bias with these metrics. After giving a theoretical explanation for the issue, we present an alternative procedure that attempts to address the problem with the use of marginal models. We then analyze the effectiveness of this procedure, both by running simulations and by applying it to actual student data. The results indicate that the marginal model procedure seemingly compensates for the bias observed in other transition metrics, thus resulting in more accurate estimates of the significance of transitions between states.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; • **Applied computing** → **E-learning**.

KEYWORDS

sequential data, transition metrics, marginal models, L statistic, affect dynamics

ACM Reference Format:
Jeffrey Matayoshi and Shamya Karumbaiah. 2021. Using Marginal Models to Adjust for Statistical Bias in the Analysis of State Transitions. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3448139.3448182

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
LAK21, April 12–16, 2021, Irvine, CA, USA
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448182

1 INTRODUCTION

As learning is a process that occurs over time, many areas of education and learning analytics research require the analysis of data that have a sequential or temporal ordering. Such analyses are important, as our understanding of the learning process can be greatly improved by leveraging the temporal features of these data [17]. Additionally, properly analyzing the sequential properties of educational data has been shown to help improve the performance and accuracy of student models [2, 19]. When dealing with sequential data, researchers are often interested in the transitions that occur between different states—or events—in these sequences. One prominent example is in the area of affect dynamics, with the goal being to identify transitions between affective states that are highly significant [10, 15]. Other works have applied similar analyses to logs of student actions in learning systems, in an attempt to understand how students transition between different activities within these sytems [6, 7].

Recently, the work in [8] evaluated several metrics commonly used to analyze transitions within sequential data. In addition to looking at the probability estimates of different transitions, the study also evaluated techniques and transition metrics such as lag sequential analysis [24] and the L statistic [11]. Using numerical simulations, the analysis revealed a subtle statisical bias that occurs with these transition metrics, causing them to return unexpected and inflated values. This bias then creates extra difficulties when interpreting the values of the transition metrics, thus making it harder to measure the significance of transitions; additionally, the experiments in [8] showed that this issue is especially pronounced in short sequences of transitions.

Motivated by these results, in this current work we look in more detail at these issues. After first replicating the numerical experiments from [8], we next give a theoretical analysis that attempts to explain the underlying mechanisms causing the observed statistical bias. Based on this explanation, we then outline a regression procedure that measures the significance of transitions using a marginal model approach. To evaluate the effectiveness of this procedure, we apply it to the simulated data generated from our numerical experiments, as well as to actual student data from studies of affect dynamics.

2 TRANSITION METRICS AND STATISTICAL BIAS

2.1 Transition Metric Simulations

Consider the case when transitions between states happen purely at chance; that is, at all times in a sequence of states, the next state is sampled uniformly at random from all possible states. In

such a case, we want our transition metric to return a baseline value that indicates the transitions are happening randomly and are not influenced by the starting state. This is the setting for the numerical experiments in [8], and we begin our current analysis with a replication of their work. To that end, consider two possible states, A and B. For each sequence length from 3 to 150, we generate 10,000 sequences of the given length by choosing between A and B at random; that is, we randomly choose between A and B with an equal probability of 0.5. For each of these sequences we compute the values of $P(B \mid A)$ and $P(A \mid A)$; the former is the probability of transitioning to B, given that the starting state is A, while the latter is the probability of transitioning to A, given that the starting state is also A. Once we've computed these values for each sequence, we then compute the average for each conditional probability over the entire group of 10,000 sequences. Additionally, as another point of comparison, as done in [8] we also include the values from the L statistic, a popular transition metric used in the field of affective dynamics. The L statistic, which was originally introduced in [11], is defined as follows.

Definition 1 (L statistic)**.** For states A and B, let $A \longrightarrow B$ represent transitions that start in state A and end in state B. We then have

$$L(A \longrightarrow B) := \frac{P(B \mid A) - P(B)}{1 - P(B)}, \tag{2.1}$$

where $P(B)$ is the overall probability of B occurring as the next state and $P(B \mid A)$ is the conditional probability of transitioning to B, given that the starting state is A.

As with the conditional probabilities, we first compute the L values individually for each sequence, and we then find the averages of these values from the entire group of sequences. The values for both the conditional probabilities and the L statistic are shown in Figure 1, and it's worth noting that the results are consistent with those from [8]. Regarding the conditional probabilities, while we expect these to be close to 0.5, as we are choosing between A and B equally at random, we can see that the computed values are heavily biased for the shortest sequences, with the bias then decreasing—but not completely disappearing—as the sequence length grows. In particular, the conditional probability values measuring transitions from $A \longrightarrow A$ are biased in the negative direction, while the values for transitions of the form $A \longrightarrow B$ show a bias in the positive direction. Turning next to the L statistic, we expect the values to be close to zero as, again, the states are being chosen uniformly at random. However, as with the conditional probabilities, we can see that there is a bias that is especially pronounced for the shortest sequences. For example, the maximum value of $L(A \longrightarrow B)$ is just over 0.4 and the minimum value of $L(A \longrightarrow A)$ is just under -0.5, and both of these values occur with the sequences of length 3. Note that, while the bias is fairly minimal once we reach sequences lengths of 40 or 50, obtaining this amount of data in a physical classroom can be challenging and impractical.[1]

[1] As a simple example, consider a relatively small classroom containing 10 students. If we assume that the observation window is 20 seconds—which is fairly standard in affect dynamics research—it would take a single observer more than 2 hours to obtain 40 observations for each student, something that isn't possible when a class period is under an hour, as many are.

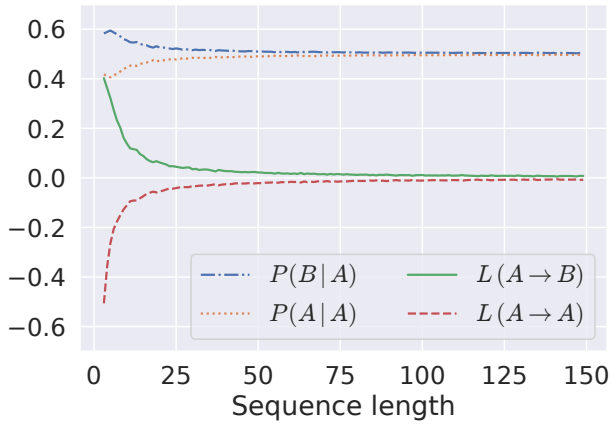


Figure 1: Plot of L values and conditional probabilities from our replication of the simulations in [8]. The sequences are generated by choosing states A and B equally at random.

2.2 Bias with Transition Metrics—Theoretical Analysis

Now that we have replicated the experiments from [8], we next offer a theoretical explanation for the biased values observed in these experiments. The core of the issue can be traced to the conditional probability estimates of $P(B \mid A)$ and $P(A \mid A)$. Our claim is that the bias is an artifact of the averaging procedure used to estimate these values across the different sequences, and that this bias is then carried through to various transition metrics, such as L, that rely on these estimates. We illustrate the issue using a simple example. Consider the eight distinct sequences of length three consisting only of the states A or B, or both.

$$AAA \quad AAB \quad ABA \quad ABB \quad BAA \quad BAB \quad BBA \quad BBB$$

Now, as all transitions are equally likely in this set of sequences—i.e., all transitions occur with the same frequency—we would expect the computed estimates of $P(B \mid A)$ and $P(A \mid A)$ to each be 0.5. However, as shown by the values in the Unweighted column of Table 1 this is not the case. If we compute the probabilities individually for each sequence, and we then compute the averages over all the sequences, we obtain a value of 0.42 for $P(A \mid A)$ and a value of 0.58 for $P(B \mid A)$. In this example, the averaging procedure ignores the number of transitions that occur within each sequence, which then distorts the estimates. For example, the sequence AAA contains two transitions that start in A, while the sequence BAB contains only one; however, this discrepancy is ignored when computing the values in the Unweighted column of Table 1. Based on these results, this effect can be summarized, in some sense, by saying that high values of $P(A \mid B)$ occur more frequently than high values of $P(A \mid A)$ when the number of transitions within the sequences are ignored.

Next, consider what happens if, instead of averaging the conditional probabilities over the sequences, we compute the conditional

probabilities by combining—or pooling—all of the data. That is, rather than grouping the transitions by sequence, we simply compute the rates of the transitions over the entire data set. Equivalently, we can also think of this as computing a weighted average of the conditional probabilities per sequence, where the weight is determined by the number of relevant transitions. For example, since sequences such as *AAA* and *AAB* contain two transitions that start in *A*, we assign these a weight of 2; on the other hand, sequences such as *ABB* and *BAB* only contain one transition from *A*, so these sequences are assigned a weight of 1. The results are shown in the Weighted column of Table 1, where we can see that the weighted conditional probabilities are both equal to 0.5, as desired.

3 REGRESSION PROCEDURE USING MARGINAL MODELS

Based on the discussion in the previous section, the bias in the conditional probability estimates can be removed by using the extra information that is lost when averaging the values for each sequence. Thus, in what follows we describe a procedure that attempts to retain this information with the use of a logistic regression model. To begin, suppose we are interested in studying transitions of the form $A \to B$. Furthermore, assume that there are no restrictions on transitions between states.[2] To estimate the effect that starting in A has on transitions to B, we build a regression model in which the response variable is binary, with a value of one if the next state is equal to B, and a value of zero otherwise. Due to the binary form of the response variable, we use the logit as our link function. The sole predictor variable is another binary variable that is one if the previous state is equal to A, and zero otherwise. Under this formulation, a sequence of length n generates $n-1$ data points. The variables of the model are summarized as follows.

- $y = y_{it}$: one if B is the next state for student i at time t; zero otherwise
- $x = x_{it}$: one if A is the previous state for student i at time t; zero otherwise

Letting σ represent the standard logistic function, the regression equation then has the form

$$P(y_{it} = 1 \mid x_{it}) = \sigma(\beta_0 + \beta_1 x_{it}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{it})}}. \tag{3.1}$$

In our analysis, we are interested in the coefficient β_1 from (3.1). The value of β_1 gives an indication of how much starting in state A—in comparison to not starting in state A—influences the likelihood of transitioning to state B. A large positive value for this coefficient would suggest that starting in A *increases* the likelihood of transitioning to B, in comparison to starting in a state that is not A. Analogously, a negative value for the coefficient might suggest that starting in A *decreases* the likelihood of transitioning to B, in comparison to not starting in A. Additionally, a benefit of this approach is that we can compare the probability estimates from the regression for the two different values of the predictor variable—zero or one—to get an additional measure of how large of an effect the predictor variable has.

[2]In comparison, some studies specifically ignore self-transitions—i.e., data points in which the same state appears consecutively—while in other situations certain transitions may be impossible. We return to this topic in the discussion.

One complication with the above procedure is that we have to deal with dependent—or correlated—data, as each sequence of transitions contains multiple measurements from the same student. Thus, when fitting the parameters for the logistic regression, we need to properly account for the dependence between these repeated measurements. We can accomplish this by using a multilevel model, where each individual student is considered a "group" or "cluster". Specifically, we use a marginal—or population averaged—model based on generalized estimating equations (GEE) [13, 18]. Marginal models are able to handle correlated data, and as such they are commonly used on data containing repeated measurements. We choose a marginal model because of our focus on estimating the average response over the entire population, rather than estimating the effects on the individuals.[3] In order to account for the correlated data, we must specify the type of correlation structure for the data within each group. In our situation with repeated measurements, two common choices for the structure are an exchangeable correlation and a first-order autoregressive correlation. The exchangeable structure assumes that there is some common dependence between all the data in a group, while the autoregressive structure assumes that the dependence between the data in a group varies with time [12, 13, 27]. While it may occasionally be difficult to precisely determine the correct choice of correlation structure, it's worth noting that the parameter estimates are statistically consistent even if this structure is misspecified; in such a case, only the efficiency of these estimates is compromised [12, 18].

Since the estimating equations used in GEE models are not necessarily likelihood based, we are unable to use the standard Akaike Information Criterion (AIC) [1] to compare different models. Instead, we can compare the fits of different models using the Quasi-AIC (QIC) score [23]; among other things, using the QIC score can help us determine the best choice of correlation structure [12, 23]. Then, to analyze the effect of our predictor variable, we can evaluate β_1 using standard techniques such as the Wald test for statistical significance [12]. Another advantage of this approach is that it directly compares the cases when (a) the starting state is A and (b) the starting state is not A. In comparison, the L statistic compares the cases when the starting state is A to the overall behavior, regardless of the starting state. The drawback to the latter approach is that if A is very common and its occurrence dominates the sequence of states, it's possible that the values of $P(B)$ and $P(B \mid A)$ will be very close simply because A is almost always the starting state.

4 EXPERIMENTS ON SIMULATED DATA

In this section we apply the marginal model approach to simulated data using the `statsmodels` [25] Python library.[4] We begin by applying the model to the data from our replication of the work in [8]; recall that, for $n = 3, 4, \ldots, 150$, we generate 10,000 different sequences, each of length n, where each state in each sequence is chosen uniformly at random from A or B. The first set of results for the transitions $A \to A$ and $A \to B$ are shown in Figure 2a. There, we plot the unweighted conditional probability values, computed

[3]If the focus is on the individuals, one possible approach is to estimate the subject-specific parameters by using a mixed-effects model with a random intercept for each student.

[4]A Python module for running these numerical experiments is available at https://github.com/jmatayoshi/sequence-analysis.

Table 1: Computed weighted conditional probabilities.

	AAA	AAB	ABA	ABB	BAA	BAB	BBA	BBB	Mean	
									Unweighted	Weighted
$P(A \mid A)$	1	0.5	0	0	1	0	–	–	0.42	0.5
$P(B \mid A)$	0	0.5	1	1	0	1	–	–	0.58	0.5
Weight	2	2	1	1	1	1	0	0		

directly from the raw data and averaged over each set of 10,000 trials, along with the estimated probabilities from the marginal models; in the latter case—i.e., the dashed green line and solid red line—these estimates correspond to the model predictions when $x = 1$. We can see that, for both transition pairs $A \rightarrow A$ and $A \rightarrow B$, the estimates from the marginal models are all closely centered around 0.5; this is in sharp contrast to the computed conditional probability values, which exhibit the previously discussed bias.

Next, in Figure 2b we compare the L values with the values of β_1, the coefficient of our single predictor variable. As shown previously, the L values for $A \rightarrow B$ have a positive bias, with a maximum value of just over 0.4, while the L values for $A \rightarrow A$ have a negative bias, with a minimum value of just below −0.5. However, in all cases the β_1 values are closely centered around zero, as is preferred. We should also mention that, for this analysis, we use the exchangeable correlation structure for the marginal models. As the states are chosen with equal probability from either A or B, there is actually no underlying dependence in the data; thus, it is instructive that, even with the incorrect correlation structure, the resulting parameter estimates are accurate.

To investigate the situation when the transition states occur with different frequencies—or base rates—we run one additional set of simulations. For these simulations, we assume there are four possible states: A, B, C, and D. To generate our sequences, we sample randomly according to the following distribution: A is chosen with probability 0.6, B is chosen with probability 0.2, and C and D are each chosen with probability 0.1. Then, for $n = 3, 4, \ldots, 150$, we generate 10,000 different sequences, each of length n, according to this probability distribution on the states. The results are shown in Figure 3, where we plot the computed conditional probabilities, along with the estimated probabilities from the marginal models. As before, we can see that the raw conditional probabilities are biased for the shorter sequences. In comparison, the estimates from the marginal models are centered closely around the true values.

5 APPLICATION TO REAL STUDENT DATA

5.1 Affect Dynamics Data Sets

Our next analysis evaluates the performance of the marginal model procedure on actual student data. Specifically, we apply the technique to two different data sets consisting of affect sequences. Our first data set comes from students working in the Physics Playground learning environment [26]. The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) [22] was used to record the affective states of 179 high school students working within this environment [3]. For our purposes, we are interested in the states flow (FLO), confusion (CON), frustration (FRU), and boredom (BOR); the remaining states have all been merged into the dummy state NA. The recorded sequences for these students are relatively long, with the mean and median lengths being 135.2 and 126.0, respectively, with a standard deviation of 68.9; the minimum sequence length is 47, while the maximum is 272.

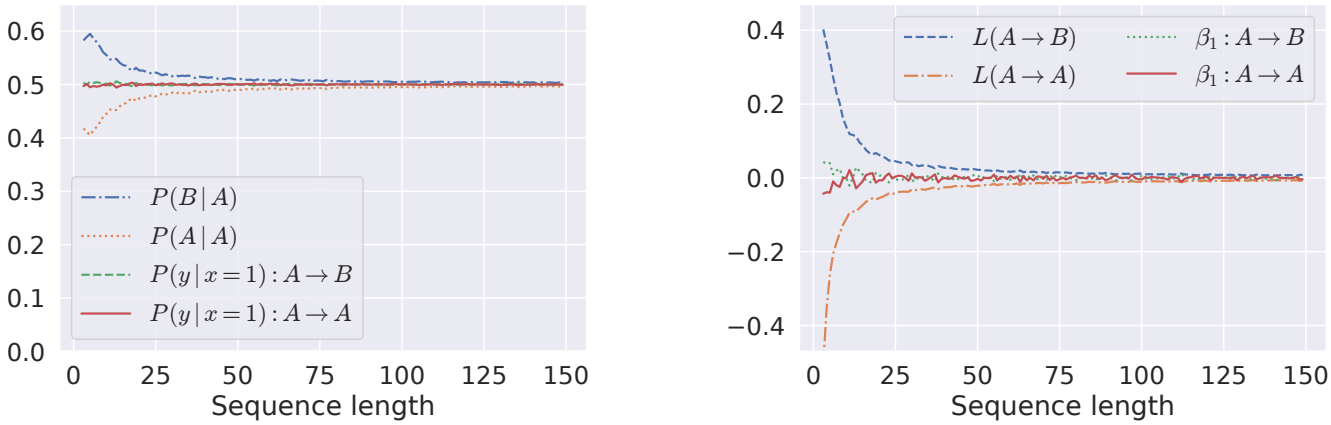
In contrast to the Physics Playground data, our second data set has very different characteristics. Namely, the sequences are much shorter, which makes for an interesting analysis, as we can see how the biases that have been observed in the simulated data affect the results from actual student data. This particular data set consists of sequences from 782 students working in the ASSISTments platform [14], with BROMP again being used to record the student affective states [9]; as before, we focus on the states flow (FLO), confusion (CON), frustration (FRU), and boredom (BOR), with any remaining states being merged into the dummy state NA. The mean and median lengths of the sequences in this data set are 9.6 and 9.0, respectively, with a standard deviation of 5.2. The minimum sequence length is 3, while the maximum is 37.

5.2 Experimental Results

Based on the experiments from [8], as well as our results in Section 4, we expect the relatively long sequence lengths in the Physics Playground data to minimize the bias in the L statistic values. In comparison, we expect to see some evidence of this bias in the ASSISTments data, due to the very short sequence lengths. The results from applying the marginal model, as well as the corresponding L values, are shown in Table 2. For this analysis we use an exchangeable correlation structure as, overall, it gives better performance in comparison to the autoregressive structure.[5]

To adjust for the number of statistical tests being performed, we have highlighted—in bold—the transition pairs that are statistically significant after applying the Benjamini-Yekutieli procedure [5], using an α value of 0.05. While the Benjamini-Hochberg procedure [4] has previously been used in the study of affect transitions [21], it has only been proven to be valid if the statistical tests are independent of each other, or under certain dependency conditions between the tests [5]. At the moment, we are not aware of any studies or theoretical results that indicate our experimental setup satisfies the requirements for applying the Benjamini-Hochberg procedure. Thus, we instead use the Benjamini-Yekutieli procedure, as it can be applied under arbitrary dependence conditions between the statistical tests [5]; furthermore, in light of several recent studies that call into question some results from previous applications

[5]While many applications of the autoregressive structure deal with time scales on the order of weeks, months, or years—e.g., epidemiological studies—the time scales for our data sets are much smaller, on the order of minutes or hours. Thus, due to these small time scales it's plausible that the dependence in our data is relatively constant over time, thereby making the exchangeable structure a better fit.



(a) (b)

Figure 2: Plots comparing (a) the unweighted conditional probability values and estimates from the marginal models, and (b) the L values and β_1 coefficients from the marginal models. The sequences are generated by choosing states A and B equally at random.

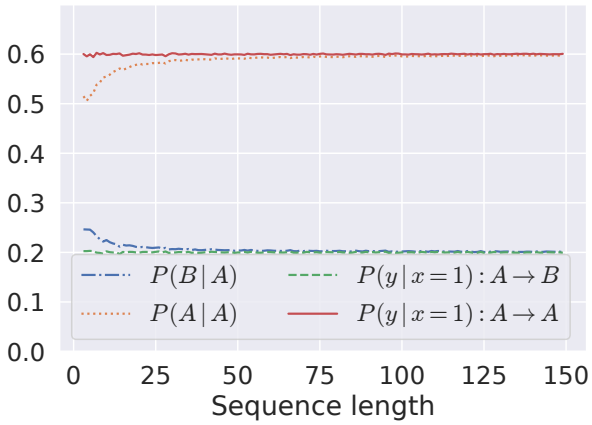


Figure 3: Plot of unweighted conditional probability values and the corresponding estimates from the marginal models. The sequences are generated by choosing A with probability 0.6, B with probability 0.2, and C and D each with probability 0.1.

of the L statistic [8, 15, 16], we believe that controlling for false discoveries with the more conservative Benjamini-Yekutieli procedure is justified.[6] Finally, note that the focus in this analysis is more on comparing the marginal model and L statistic results, and less about identifying and interpreting significant affect transitions. To that end, and to more faithfully simulate the results from an actual study of affect transitions, we perform the Benjamini-Yekutieli correction twice: once for all the marginal model values, and then separately for all the L statistic values.

Starting with the results from the Physics Playground data, it appears that the longer sequences in this data set have mitigated the effects of the bias with the L values. That is, there are only four transition pairs in the Physics Playground data for which the sign of the L statistic differs from the sign of the corresponding β_1 value; in all of these examples, the associated p-values are relatively high (0.16 or larger), indicating a lack of strong evidence that the L and β_1 values are different from zero. Furthermore, all of the self-transitions values, for both β_1 and L, are positive, with $p \ll 0.001$ in all cases; note that while the bias with the L statistic can cause the importance of self-transitions to be underestimated, this does not appear to be the case here, most likely because of the long sequences of transitions.

Next, looking at the results for the ASSISTments data, there appears to be evidence of the bias in the L values for these shorter sequences. For example, recall that on short sequences of simulated data, the L statistic returns considerably lower than expected values for self-transitions. Thus, it is instructive to see that in all self-transition cases the L values are negative and significantly different from zero, while all the corresponding β_1 values are positive and significantly different from zero. For comparison, of the 20 transitions between different states, there are only 4 transition pairs that have negative L values, none of which are significantly different from zero; thus the positive bias when the L statistic is applied to transitions between different states seems to be a factor. Furthermore, in all four of these cases with negative L values, the corresponding β_1 values are also negative, possibly indicating that

[6] Motivated by these issues, we are in the middle of a more thorough evaluation of the appropriateness of applying the Benjamini-Hochberg procedure when analyzing sequential data.

prev	*next*	Physics Playground (longer sequences)						ASSISTments (shorter sequences)					
		Marginal model				L statistic		Marginal model				L statistic	
		β_1	p-value	$x=0$	$x=1$	Mean	p-value	β_1	p-value	$x=0$	$x=1$	Mean	p-value
FLO	FLO	**0.78**	0.000	0.61	0.78	**0.13**	0.000	**0.51**	0.000	0.57	0.69	**-0.09**	0.000
	CON	**-0.47**	0.000	0.08	0.05	**-0.01**	0.000	**-0.35**	0.002	0.08	0.06	**0.02**	0.001
	FRU	**-0.63**	0.000	0.08	0.04	**-0.01**	0.000	-0.21	0.170	0.04	0.03	0.00	0.285
	BOR	**-1.43**	0.000	0.06	0.02	**-0.02**	0.000	**-0.38**	0.000	0.12	0.08	**0.03**	0.002
	NA	**-0.37**	0.000	0.15	0.11	**-0.01**	0.000	**-0.61**	0.000	0.22	0.13	-0.01	0.293
CON	FLO	**-0.62**	0.000	0.74	0.60	**-0.71**	0.00	-0.02	0.870	0.64	0.64	**0.18**	0.004
	CON	**1.33**	0.000	0.05	0.18	**0.09**	0.000	**0.77**	0.000	0.06	0.12	**-0.12**	0.000
	FRU	**0.38**	0.000	0.05	0.07	0.04	0.013	0.35	0.270	0.03	0.04	0.00	0.797
	BOR	-0.78	0.062	0.03	0.01	-0.02	0.014	-0.12	0.500	0.10	0.09	-0.02	0.339
	NA	-0.06	0.462	0.13	0.12	0.01	0.700	-0.02	0.878	0.16	0.16	0.03	0.394
FRU	FLO	**-0.81**	0.000	0.74	0.56	**-0.42**	0.00	**-0.48**	0.000	0.64	0.53	-0.07	0.401
	CON	0.14	0.358	0.06	0.07	-0.00	0.983	0.48	0.068	0.06	0.10	0.04	0.163
	FRU	**1.60**	0.000	0.04	0.18	**0.07**	0.000	-0.01	0.988	0.03	0.03	**-0.10**	0.000
	BOR	0.38	0.034	0.03	0.04	0.03	0.018	0.52	0.011	0.10	0.15	0.05	0.233
	NA	-0.00	0.979	0.12	0.12	0.03	0.163	0.40	0.019	0.16	0.22	0.09	0.032
BOR	FLO	**-1.28**	0.000	0.74	0.44	**-0.77**	0.00	**-0.28**	0.003	0.65	0.58	**0.18**	0.001
	CON	**-1.16**	0.000	0.06	0.02	**-0.05**	0.000	-0.31	0.136	0.06	0.05	-0.01	0.314
	FRU	-0.17	0.279	0.05	0.04	0.01	0.390	0.31	0.229	0.03	0.04	0.01	0.280
	BOR	**2.96**	0.000	0.02	0.27	**0.23**	0.000	**0.73**	0.000	0.09	0.16	**-0.09**	0.000
	NA	-0.25	0.126	0.13	0.10	-0.02	0.400	0.07	0.553	0.16	0.17	0.01	0.625
NA	FLO	**-0.17**	0.007	0.73	0.70	-0.01	0.804	**-0.37**	0.000	0.65	0.57	0.11	0.033
	CON	**-0.25**	0.003	0.06	0.05	-0.01	0.076	0.07	0.595	0.06	0.07	0.01	0.434
	FRU	**-0.47**	0.000	0.05	0.03	-0.02	0.010	-0.06	0.779	0.03	0.03	0.00	0.771
	BOR	**-0.78**	0.003	0.03	0.02	**-0.02**	0.001	-0.07	0.527	0.10	0.09	0.02	0.122
	NA	**0.65**	0.000	0.11	0.20	**0.06**	0.000	**0.79**	0.000	0.14	0.26	**-0.05**	0.002

Table 2: Comparison of the marginal model and L statistic results on the Physics Playground [3] and ASSISTments [9] data sets. Bold values are statistically significant after applying the Benjamini-Yekutieli procedure with an α value of 0.05. Note that, when applying the Benjamini-Yekutieli procedure, we have applied it separately for the marginal model values and the L values, in order to simulate the workflow in a typical study of affect transistions.

the negative relationships between the states are strong enough to overcome the positive bias with the L statistic. Thus, the overall results on the ASSISTments data set are seemingly consistent with the biases that appear in the experiments on simulated data.

6 DISCUSSION

As shown by the work in [8], several commonly used transition metrics suffer from a bias that can inflate the significance of transition pairs. Motivated by these concerns, in this work we presented the results from further investigation of these issues. We began this analysis by replicating the numerical experiments in [8]. Next, we presented a theoretical explanation for the underlying cause of this bias, where we argued that it's a consequence of the averaging procedure commonly used in the computations of transition metrics. Based on these results, we then outlined a procedure for measuring the importance and significance of transition pairs. This procedure takes the form of a logistic regression that estimates the probability of transitioning to a state, depending on the occurence of a previous state; the parameters for the regression were obtained using marginal models. To show that this procedure does not suffer from the bias inherent in other transition metrics, we examined its effectiveness on both simulated and real data.

Note that the approach we outlined here is flexible, as it can be applied to estimate and measure the effects of other relationships beyond a single transition between states. For example, suppose we are interested in whether starting in state A has an influence on the appearance of the sequence BAB as the next three states. In this case, we simply need to change our response variable to fit the situation. Rather than defining the response variable based on the next state, we simply change the definition so that it has a value of one if the next three states are BAB, and a value of zero otherwise. Or, perhaps we are still interested in a transition to a single state, but rather than looking at the starting states individually, we would rather directly compare their influence simultaneously. In this case, we can use different indicator variables for the starting states, and then compare the coefficients of these indicator variables to get a relative ordering of significance of the different starting states.

Lastly, another potential application occurs in situations where transitions between certain states are excluded, either by necessity or by the preference of the researcher. For example, several recent works have looked at cases when self-transitions—i.e., examples

where the same state appears consecutively—are excluded from sequences of student affect data, and these analyses have demonstrated that issues arise when the L statistic is applied in these situations [8, 15, 16, 20]. At the moment, we believe that the regression procedure described in this work can be adapted to these cases, and as such we are currently in the process of evaluating the validity of this approach.

REFERENCES

[1] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723.

[2] Alejandro Andrade, J. Danish, and Adam Maltese. 2017. A Measurement Model of Gestures in an Embodied Learning Environment: Accounting for Temporal Dependencies. *Journal of learning Analytics* 4 (2017), 18–46.

[3] Juan Miguel L. Andres, Ma Mercedes T. Rodrigo, Jessica O. Sugay, Michelle P. Banawan, Yancy Vance M. Paredes, Josephine S. Dela Cruz, and Thelma D. Palaoag. 2015. More Fun in the Philippines? Factors Affecting Transfer of Western Field Methods to One Developing World Context.. In *AIED Workshops*.

[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.

[5] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* (2001), 1165–1188.

[6] Gautam Biswas, Hogyeong Jeong, J. Kinnebrew, Brian Sulcer, and Rod D. Roscoe. 2010. Measuring Self-Regulated Learning Skills through Social Interactions in a teachable Agent Environment. *Res. Pract. Technol. Enhanc. Learn.* 5 (2010), 123–152.

[7] Nigel Bosch and Sidney D'Mello. 2017. The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education* 27, 1 (2017), 181–206.

[8] Nigel Bosch and Luc Paquette. 2019. What's Next? Edge Cases in Measuring Transitions Between Sequential States. (2019). Submitted.

[9] Anthony F. Botelho, Ryan S. Baker, and Neil T. Heffernan. 2017. Improving Sensor-Free Affect Detection Using Deep Learning. In *Artificial Intelligence in Education-18th International Conference, AIED 2017*. 40–51.

[10] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[11] Sidney D'Mello, Roger S. Taylor, and Art Graesser. 2007. Monitoring Affective Trajectories During Complex Learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 29 (29)*. 203–208.

[12] James W. Hardin and Joseph M. Hilbe. 2012. *Generalized Estimating Equations*. Chapman and Hall/CRC.

[13] Patrick J. Heagerty and Scott L. Zeger. 2000. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statist. Sci.* 15, 1 (2000), 1–26.

[14] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.

[15] Shamya Karumbaiah, JMAL Andres, Anthony F. Botelho, Ryan S. Baker, and Jaclyn S. Ocumpaugh. 2018. The Implications of a Subtle Difference in the Calculation of Affect Dynamics. In *Proceedings of the 26th International Conference on Computers in Education*. 29–38.

[16] Shamya Karumbaiah, Ryan S. Baker, and Jaclyn Ocumpaugh. 2019. The Case of Self-transitions in Affective Dynamics. In *Artificial Intelligence in Education-20th International Conference, AIED 2019*. 172–181.

[17] Simon Knight, Alyssa Friend Wise, and Bodong Chen. 2017. Time for Change: Why Learning Analytics Needs Temporal Analysis. *Journal of Learning Analytics* 4, 3 (2017), 7–17.

[18] Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 1 (1986), 13–22.

[19] Mohammad Javad Mahzoon, Mary Lou Maher, Omar Eltayeby, Wenwen Dou, and Kazjon Grace. 2018. A Sequence Data Model for Analyzing Temporal Patterns of Student Data. *Journal of Learning Analytics* 5, 1 (2018), 55–74.

[20] Jeffrey Matayoshi and Shamya Karumbaiah. 2020. Adjusting the L Statistic when Self-Transitions are Excluded in Affect Dynamics. *Journal of Educational Data Mining* 12, 4 (2020), 1–23.

[21] Jaclyn Ocumpaugh, Juan Miguel Andres, Ryan Baker, Jeanine DeFalco, Luc Paquette, Jonathan Rowe, Bradford Mott, James Lester, Vasiliki Georgoulas, Keith Brawner, et al. 2017. Affect dynamics in military trainees using vmedic: From engaged concentration to boredom to confusion. In *International Conference on Artificial Intelligence in Education*. Springer, 238–249.

[22] Jaclyn Ocumpaugh, Ryan S. Baker, and Ma Mercedes T. Rodrigo. 2015. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences* 60 (2015).

[23] Wei Pan. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics* 57, 1 (2001), 120–125.

[24] Gene P Sackett. 1979. The Lag Sequential Analysis of Contingency and Cyclicity in Behavioral Interaction Research. *Handbook of Infant Development* 1 (1979), 623–649.

[25] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

[26] Valerie Jean Shute and Matthew Ventura. 2013. *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

[27] Camille Szmaragd, Paul Clarke, and Fiona Steele. 2013. Subject specific and population average models for binary longitudinal data: a tutorial. *Longitudinal and Life Course Studies* 4, 2 (2013), 147–165.