JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 5(2), 86–104. http://dx.doi.org/10.18608/jla.2018.52.6

Metrics for Discrete Student Models: Chance Levels, Comparisons, and Use Cases

Nigel Bosch [1], Luc Paquette [2]

Abstract

Metrics including Cohen's kappa, precision, recall, and F_1 are common measures of performance for models of discrete student states, such as a student's affect or behaviour. This study examined discrete model metrics for previously published student model examples to identify situations where metrics provided differing perspectives on model performance. Simulated models also systematically showed the effects of imbalanced class distributions in both data and predictions, in terms of the values of metrics and the chance levels (values obtained by making random predictions) for those metrics. Random chance level for F_1 was also established and evaluated. Results for example student models showed that over-prediction of the class of interest (positive class) was relatively common. Chance-level F_1 was inflated by over-prediction; conversely, maximum possible values for F_1 and kappa were negatively impacted by over-prediction of the positive class. Additionally, normalization methods for F_1 relative to chance are discussed and compared to kappa, demonstrating an equivalence between kappa and normalized F_1. Finally, implications of results for choice of metrics are discussed in the context of common student modelling goals, such as avoiding false negatives for student states that are negatively related to learning.

Notes for Practice

- Previous research has shown that choice of metric plays a key role in training and evaluation of student models, focusing primarily on metrics intended for models that produce probabilistic predictions of student outcome variables
- Imbalances in labelled data are quite common in student modelling tasks, and have been shown to impact metrics used for machine-learned student models
- This paper explores the impact that predicted class proportions and data class proportions have on discrete model metrics including Cohen's kappa, precision, recall, and F_1, and formulates a random-chance level F_1 measurement that is adjusted for imbalances
- Results on real-world student models and simulated models show that best practices include reporting multiple metrics for discrete student models, and comparing F_1 scores to the appropriate chance level to avoid over- or under-estimating model performance

Keywords

Metrics, chance level F_1, Cohen's kappa, precision, recall, discrete student models, imbalanced data, imbalanced predictions

Submitted: 15.10.18 — **Accepted:** 06.02.18 — **Published:** 05.08.18

Corresponding author [1]Email: pnb@illinois.edu Address: National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1205 W Clark Street, Urbana, IL 61801, United States ORCID ID: 0000-0003-2736-2899

[2]Email: lpaq@illinois.edu Address: Department of Curriculum & Instruction, University of Illinois at Urbana-Champaign, 1310 S Sixth Street, Champaign, IL 61820, United States

1. Introduction

Predicting student state, actions, or outcomes (student modelling) is one of the largest and most diverse areas within learning analytics research (see Chrysafiadi & Virvou, 2013; Desmarais & Baker, 2012; Henrie, Halverson, & Graham, 2015; Papamitsiou & Economides, 2014 for recent reviews). Student models can ascertain a host of student attributes (e.g., how much does a student know about the topic they are currently studying), and can detect and

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

predict important learning-related states students may be in. These states might include a student's current emotional, cognitive, or behavioural state (Bailey & Konstan, 2006; Baker, Corbett, Koedinger, & Wagner, 2004; Baker, D'Mello, Rodrigo, & Graesser, 2010; Bixler & D'Mello, 2015; Bosch, D'Mello, Ocumpaugh, Baker, & Shute, 2016; Calvo & D'Mello, 2010; Walonoski & Heffernan, 2006). Such models are tremendously important because they allow greater scientific understanding of learning, enable the construction of more effective and enjoyable computerized learning environments, and drive better feedback for both teachers and students. Evaluating the accuracy (i.e., model performance) of such student models is thus crucial to answer questions such as whether model A is "better" than model B, or whether a model intended to intervene when a student becomes bored is better than a model that simply triggers an intervention randomly.

Assessing whether a student is bored or not is an example of a discrete modelling task. The student model must make a binary decision about whether the student is bored, or provide a probability estimate that the student is bored. Other examples include modelling student attention (Raca, Kidzinski, & Dillenbourg, 2015), off-task activity (Bosch et al., 2016), or exploitive behaviour (a.k.a. "gaming the system"; Baker et al., 2004). Such modelling tasks are common in educational environments, as a variety of student states are related to learning (Bower, 1992; Kort, Reilly, & Picard, 2001; McVay & Kane, 2009; Nissen & Bullemer, 1987; Pekrun, Goetz, Titz, & Perry, 2002; Smallwood, Fishman, & Schooler, 2007; Trigwell, Ellis, & Han, 2012). Evaluating models for these tasks typically results in a single number or a small set of numbers intended to succinctly characterize the performance of the model. These evaluation methods are typically referred to as metrics, though they are not often "metric" in the mathematical sense (Lawvere, 1973). Different metrics are required for student models that detect discrete states than those models that detect or predict continuous states and outcomes, such as final percentage grade in a course.

There are three categories of metrics for common types of student models. The first is metrics for continuous-valued (real number) predictions of continuous-valued outcome variables. For example, root mean-squared error could measure the difference between predicted and actual time spent on the next problem. The second type of metric is for models generating continuous-valued predictions for discrete (e.g., integer or binary) outcomes, such as a model that predicts the probability that a student will answer the next problem in a sequence correctly. Metrics for this type of model have been well-researched (Pelánek, 2015), and include metrics such as area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPRC). In an example application, a logistic regression model might be trained to predict the odds (a continuous prediction) of a student completing an online course (a discrete outcome) based on records of their actions in a learning environment (Robinson, Yeomans, Reich, Hulleman, & Gehlbach, 2016). In this case, this continuous prediction can be used to make a binary decision — such as whether or not to intervene for low-performing students — by choosing a threshold after model training. Choosing this threshold directly impacts the proportion of each class predicted by the model, and thus also influences the chance level (values obtained by making random predictions) of metrics discussed in this article.

The main focus of this article is a third, less well-explored category of metrics: those intended for models with discrete predictions of discrete outcomes. For example, a model to predict whether a student is bored or not (a discrete outcome) may be created with a nearest neighbour classifier. This classifier produces discrete predictions (bored/not bored) based on whether a student is most similar (in terms of facial expression, speech patterns, eye movements, etc.) to another student who is bored or not bored. Several types of models, including rule learners, such as RIPPER (Cohen, 1995), and decision trees, such as C4.5 (Quinlan, 1993), produce discrete predictions and require appropriate metrics to measure performance. The metrics we examine in particular are proportion correctly classified (commonly referred to simply as *accuracy*), Cohen's kappa, F_1 score, precision, and recall (see section 3 for definitions).

Discrete model metrics typically represent model performance as a single number. Using a single number is needed for comparing and sorting models by performance, to select the best model from a pool of candidates, or to examine trends in various models. There are pitfalls associated with reducing evaluation of a model to a single number, however. These pitfalls are primarily related to imbalances in the proportions of ground truth class labels (i.e., data class proportions) and imbalances in label predictions made by a student model (i.e., predicted class proportions) that can cause metrics to provide conflicting measures of performance (e.g., high precision but low recall, or vice versa). The effect of data class proportions has been quantified for some discrete model metrics (e.g., accuracy, Cohen's kappa, Krippendorf's alpha, F_1 score; Jeni, Cohn, & De la Torre, 2013), but the influence of predicted class proportions is less well-studied and equally complex.

One might expect that predicted class proportions should match data class proportions, and that it is thus less of a concern. There are situations where this may not be the case, however. For example, Figure 1 illustrates varying the decision threshold of a logistic regression model that predicts whether a university student is enrolled as a science major or not (Bosch et al., 2018). In this case, kappa and precision can be improved by predicting fewer positive cases (46.2%) than the data class proportions suggest (68.3%). Furthermore, in situations where false positive or false negative predictions have unequal importance, a model might favour one over the other. For example, suppose a model has been developed to predict when a student is experiencing task-unrelated thoughts (mind wandering) so that a pop quiz can be administered to measure the effect of mind wandering on retention. Mind wandering might occur during 23% (positive data proportion) of the learning session (e.g., Hutt et al., 2017), but administering quizzes during 23% of the learning session would be far too many quizzes. Thus, a good model might make only a few predictions of mind wandering, focusing on avoiding false positives but allowing for many false negatives. However, the performance of this model might appear better (or worse) than expected according to popular metrics.

Proportion correct (accuracy) is especially influenced by data class proportions and predicted class proportions, as can be seen in a simple example. Supposing a student is labelled as off task in 5% of cases in a dataset, then a model could be 95% correct by simply labelling all instances as on task (100% positive class predictions). Similarly, precision, recall, and F_1 do not correct for random chance levels that vary due to predicted and data class proportions. Recall and F_1, in particular, can be inflated by over-predicting the positive class. Nevertheless, they offer a valuable perspective into model performance, and are reported in student modelling and related literature (Cetintas, Si, Xin, & Hord, 2010; Chen, Vorvoreanu, & Madhavan, 2014; Neiberg, Elenius, & Laskowski, 2006; Pardos, Baker, San Pedro, Gowda, & Gowda, 2013; Soleymani, Pantic, & Pun, 2012; Stewart, Bosch, & D'Mello, 2017; Valstar, Mehu, Jiang, Pantic, & Scherer, 2012). Thus, understanding the chance levels for these metrics is important.

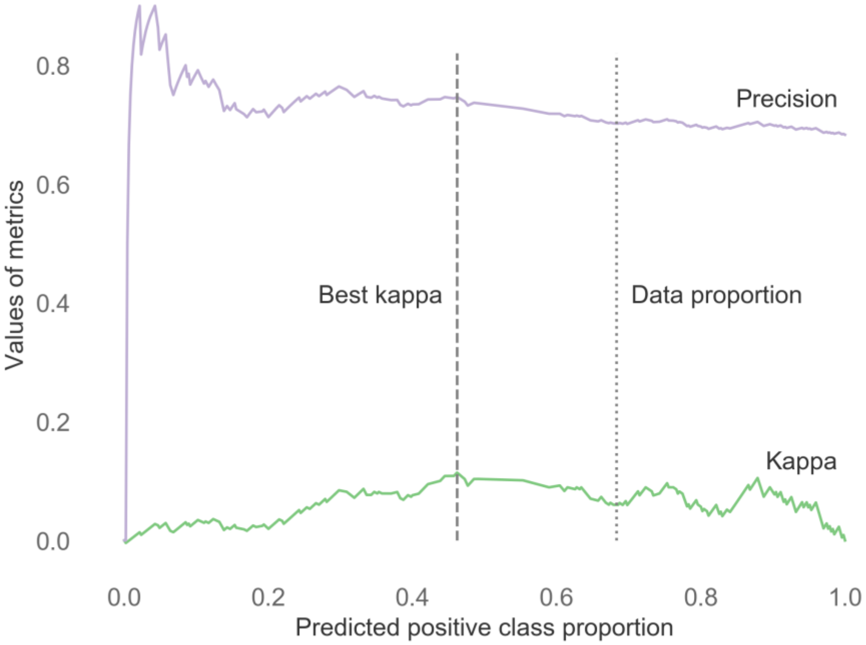

Figure 1. Precision and kappa versus positive prediction rate, illustrating a logistic regression model for which mismatching predicted and data class proportions improves accuracy.

1.1. Current Contribution

This article therefore makes two main contributions: 1) We establish and compare chance levels for these metrics where chance is not well defined (precision, recall, and F_1), and 2) We compare the effects of data class proportions and predicted class proportions on discrete model metrics (precision, recall, F_1, and kappa) with real and simulated student models, providing guidance for which metrics to use in different scenarios and how multiple metrics provide different perspectives into model performance.

For the first contribution, we consider chance levels of metrics with respect to a specific predicted class (e.g., precision of frustration predictions from a model of student affect) referred to as the positive class. In the second contribution, metrics are compared for student models on various datasets intended to illustrate the effects of data class proportions and predicted class proportions on these metrics. These experiments also reveal situations where it is preferable to consider one metric over another, as well as situations where it is necessary to consider multiple metrics to obtain a complete picture of classification accuracy.

These contributions are unique in several respects. First, this article is the first to formulate chance-level F_1 for common use cases in student modelling. We also relate chance-level F_1 to kappa, while making explicit the assumptions underlying chance level and normalization. Second, while prior research has explored the effects of data class proportions for some modelling tasks (e.g., Jeni et al., 2013), this article is the first to explicitly consider and compare the effect of predicted class proportions on metrics for discrete student models. Prior work has largely focused on models that produce continuous-valued predictions (Pelánek, 2015) or concluded that such metrics are preferable (Jeni et al., 2013). However, not all student models produce continuous predictions, and thus in-depth evaluation of discrete model metrics is also needed.

2. Related Work

We discuss research comparing metrics for model evaluation. We first review work concerning discrete metrics in general, then discuss work specifically related to student modelling (and similar tasks), and then review exemplary work on metrics for discrete models in other domains.

2.1. Research on Discrete Metrics

Jeni et al. evaluated the effect of data class proportions on various metrics (Jeni et al., 2013), including metrics for discrete predictions (accuracy, kappa, F_1, and Krippendorf's alpha), and continuous predictions (AUC and area under the precision-recall curve). They created simulated datasets and models to precisely examine the effects of data class proportions. Of the metrics considered they found all discrete metrics were affected by imbalanced data class proportions, but that AUC was not. For example, kappa was reduced by as much as ≈80% (from .6 to .1) when data imbalance was increased from a 1:1 ratio of instances of two different classes to a 50:1 ratio.

In the same paper (Jeni et al., 2013), the authors trained models on three real datasets to detect the presence or absence of facial action units, which measure the activation of facial muscles (Ekman & Friesen, 1978). Results on action unit datasets replicated their findings regarding the effect of simulated data class proportions. For example, F_1 averaged across all three action unit datasets was .28, but after sampling the datasets to balance the data class proportions, F_1 increased to .70. Upsampling, downsampling, or re-weighting *training* data is possible and even common in practice. However, such sampling is not possible in applications of models to new data where labels are not known, as is usually the case.[1] Thus, performance is typically evaluated on *testing* data with the true data class proportions. However, modifying the testing data class proportions (as in Jeni et al., 2013) does illustrate one of the dramatic effects data class proportions can have. The work of Jeni et al. is particularly relevant because they considered metrics for models that produce discrete predictions. They did not, however, report the effect of predicted class proportions (their simulated models did not have imbalanced predicted class proportions), nor did they discuss the relationship between data class proportions, predicted class proportions, and chance levels of the metrics.

Lobo, Jiménez-Valverde, and Real (2008) evaluated the AUC metric for models with probabilistic predictions. They noted several flaws that are relevant to student models. Most notably, they observed that AUC evaluates performance across all possible prediction thresholds. Model performance is evaluated where all instances except one are classified as positive, all but two classified as positive, and so on. Thus, the model is evaluated at thresholds where it would not realistically be used in a learning context (e.g., a student model that predicts students are bored

[1] For example, Hutt et al. (2017) trained a model to predict when students were mind wandering while using an intelligent tutoring system, with the goal of applying the model in real-time to new students. In this real-time application, whether students are mind wandering or not is unknown — hence the need for a predictive model. Thus, data cannot be sampled in real-time to provide equal class proportions to the model. The same applies to student models intended to predict emotion, future problem correctness, or other states and outcomes. The goal of evaluating model performance on unseen *testing* data is to measure how accurate a model is likely to be in such a real-world application. Sampling the testing data to create balanced data class proportions would thus provide an inaccurate perspective of what the model performance would be when applied (at which point the model would encounter the true data class proportions).

99% of the time). They also noted that false positive and false negative errors are considered equally. In practice, for a student model it is possible that one type of error is more serious based on context. For example, a missed instance of frustration in an affect detection model might be preferable to a missed instance of boredom because boredom can be more detrimental to learning (Baker et al., 2010). To help address these issues, Lobo et al. (2008) recommend reporting sensitivity and specificity (i.e., recall for both positive and negative classes), though this does not represent model performance as a single number.

Powers (2011) discussed the biases of discrete model metrics, including precision, recall, and F_1. Powers noted that the typical usage of precision, recall, and F_1 is to measure the performance of a model with respect to only the positive class in binary models. Importantly, this article established chance-corrected measures for precision and recall, referred to as Markedness and Informedness respectively. Chance precision was given as the base rate of the positive class as predicted by the model, while chance-level recall was given as the original base rate of the positive class in the data. These are the same precision and recall chance levels we report below. Notably, however, chance-level F_1 was not established. Using multiple metrics such as Markedness and Informedness does provide additional insight into a model, but single-value metrics such as F_1 are necessary for model ranking and selection.

2.2. Metrics for Evaluating Student Models

Pelánek (2015) evaluated a set of metrics for student models, focusing primarily on models of student knowledge such as Bayesian knowledge tracing (Baker, Corbett, & Aleven, 2008; Desmarais & Baker, 2012). Student knowledge models typically predict continuous estimates of knowledge, such as the probability that a student has mastered a specific topic, rather than discrete predictions. In Pelánek (2015), example models demonstrated flaws in some commonly reported metrics, including mean absolute error (MAE) and AUC. MAE was shown to favour models with predictions that did not reflect the true probabilities of classes, instead being biased toward models that overestimate the probability of the majority class. Instead, root mean squared error (RMSE) was recommended as a preferred metric, at least for models of student knowledge. RMSE requires continuous-valued predictions, however, so it is unsuitable for discrete predictions (e.g., from a decision tree with only five leaves). Discrete model metrics were mentioned, including accuracy, kappa, precision, recall, and F_1, but not evaluated.

Gardner and Brooks (2017) surveyed a wide range of publications that reported student models in massive open online courses. They found that accuracy, AUC, F1, precision, recall, and kappa were the most commonly reported metrics. Other than AUC, which requires continuous-valued predictions of discrete outcomes, the remaining metrics are for discrete predictions of discrete outcomes. This survey confirmed the prevalence of these metrics in student modelling research.

2.3. Discrete Models in Other Domains

F_1 is commonly used in the field of medicine to evaluate models (e.g., for detecting cancer in images or diagnosing illness). For example, Roux et al. (2013) compared the results of several different teams competing to detect mitosis in microscope images of breast cancer cells. Data class proportions were extremely imbalanced, with less than 0.3% of all instances being positive. Teams produced predictions with greatly varying predicted class proportions as well, ranging from just 28 positive predictions to 35,661, though there were only 100 positive cases in the data. Results demonstrate some of the effect of predicted class proportions on both precision and recall — the team with the most over-prediction scored the highest recall (most true positives), but also the lowest precision due to many false positives.

Hripcsak and Rothschild (2005) noted that in ground truth coding tasks where reliability needs to be assessed, F_1 approaches kappa as the proportion of negative cases grows. In modelling tasks such as mitosis detection where imbalance in data class proportions is extreme (Roux et al., 2013), chance-level F_1 approaches zero because chance-level precision also approaches zero. This is particularly interesting because, as we show below, F_1 and kappa are equivalent after controlling for F_1 chance level. Hripcsak and Rothschild's findings hint at this relationship between F_1 and kappa by comparing them in a situation where chance level for both is the same (i.e., zero).

Forman (2003) studied various feature selection methods in the domain of text classification, including methods based on optimizing discrete classification metrics. They ranked features (words of the text) according to metrics calculated by classifying a text based on that word alone, and then selected the best features. They considered F_1 as one of the metrics, noting that since it is calculated from the positive class it often leads to poor precision. They also noted that imbalanced data class proportions were related to a reduction in F_1 score. Our findings illuminate both of these phenomena in the context of student model evaluation and suggest that chance level plays a key role in similar observed patterns.

3. Method

We first briefly describe chance level for Cohen's kappa, which is an inherent part of the definition of the metric. We then explore how F_1 chance level is defined for several different scenarios, and finally describe how F_1 and other metrics discussed in the introduction are employed to evaluate student models on real and simulated datasets. Note that metrics and chance levels are presented in terms of intuitive variables such as "Number of correct positive classifications," but can also be described in terms of the cells of a confusion matrix (see Table 1 for an example): true positives, true negatives, false positives, and false negatives. For example, "Number of positive classifications" is equal to true positives + false positives.

3.1. Kappa Chance Level

It is important to understand the chance level of a metric to evaluate the extent to which a model makes classifications better than a random baseline. Kappa corrects for predicted class proportions by subtracting the accuracy of a model that makes random predictions at the same rate as the model being measured (same predicted class proportions). The resulting value is then normalized so that it is expressed as a proportion of the possible improvement over chance level (equation 1).

$$kappa = \frac{Accuracy - Chance\ accuracy}{1 - Chance\ accuracy} \tag{1}$$

The definition of random chance accuracy employed in calculating kappa and the choice of normalization illustrate one way in which metrics for discrete student models, like accuracy, can be interpreted relative to chance levels.

3.2. Formulation of F_1 Chance Level

The F_1 metric (equation 4) consists of the harmonic mean of precision (equation 2) and recall (equation 3), and thus chance-level F_1 can be formulated from the chance levels of these individual components. We describe two scenarios that occur when modelling students, and how precision, recall, and F_1 chance levels relate to these.

$$precision = \frac{Number\ of\ correct\ positive\ classifications}{Number\ of\ positive\ classifications} \tag{2}$$

$$recall = \frac{Number\ of\ correct\ positive\ classifications}{Number\ of\ positive\ instances\ in\ the\ data} \tag{3}$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

3.2.1. F_1 Chance Level with Known Data and Predicted Class Proportions

First, we consider the case in which the proportion of the positive class in a particular dataset is known, as well as the proportion of the dataset that the model predicts as the positive class. This is the most common scenario in student modelling, especially for machine-learned models where labelled training and testing data clearly indicate the data class proportions. A model in this scenario will produce labels when applied to testing data, which thus estimate the predicted class proportions. The question of random chance level can then be defined as "What F_1 would result if a model made classification decisions with a random number generator, with the same number of predictions of each class as the model being compared against?" An equivalent way of formulating this baseline is to take the predictions of a model and randomly shuffle them, so that the same predicted class proportions are preserved (same number of predictions of each class) but the associations between predicted and true labels are

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

randomized.

The chance-level recall is equal to the predicted positive class proportion because each positive instance in the data has a *predicted positive class proportion* probability of being labelled correctly as positive. For example, consider a dataset with 80 positive instances, 20 negative instances (80% data positive class proportion), and a student model with predicted positive class proportion of 70%. For any one of the 80 positive instances, there is a 70% chance the instance will be classified correctly as positive due to the predicted class proportions. Thus, 80 × 70% = 56 instances will be classified correctly, and so recall is 56 ÷ 80 = 70%.

Similarly, chance-level precision is equal to the data's positive class proportion because each positive prediction made by the model has a *data positive class proportion* probability of truly being a positive instance. In the same example as above, the data positive class proportion is 80%, and so any positive prediction randomly made has an 80% chance of being a correct positive prediction. If the model makes 70 positive predictions, 70 × 80% = 56 positive instances will be correctly classified, and thus precision is 56 ÷ 70 = 80%.

Knowing chance-level precision and recall, chance-level F_1 can be easily constructed (equation 4) by substituting chance levels for precision and recall. In the previous example, chance-level F_1 is therefore 2 × .80 × .70 ÷ (.80 + .70) = .747. The student model in question should thus exceed this number to indicate above-chance performance. Comparing F_1 to a naïve baseline without considering imbalances in data and predicted class proportions can result in an inaccurate picture of performance. In the above example, ignoring the effect of predicted class proportions on F_1 would allow a randomized classifier to appear above chance by inflating recall. A model operating at the random-chance level F_1 of .747 might be mistaken for being superior to a naïve baseline (F_1 = .700) that assumes no imbalance in predicted class proportions.

This raises a question about how F_1 and its corresponding chance level should be compared, and perhaps combined into a single number. Kappa (equation 1) subtracts chance accuracy from observed accuracy and normalizes the result on the [–1, 1] interval, which may be restricted if predicted class proportions are imbalanced (Figure 4). If the same approach is applied to F_1 (i.e., $normalized\ F_1 = (F_1 - chance\ F_1)/(1 - chance\ F_1)$), the resulting expression is equivalent to kappa. That is,

$$\frac{F_1 - chance\ F_1}{1 - chance\ F_1} = kappa = \frac{accuracy - chance\ accuracy}{1 - chance\ accuracy} \tag{5}$$

This equivalence is easily verified by expressing F_1, chance F_1, and kappa in terms of confusion matrix variables (true and false positives, true and false negatives) and simplifying the equivalence in equation 5 with a computer algebra system (e.g., the *Solve* function in Wolfram Mathematica). Clearly there is a close connection between the two metrics. Even though F_1 is not equivalent to accuracy alone, when accuracy and F_1 are adjusted for their respective chance levels the result is the same.

3.2.2. F_1 Chance Level with Unknown Data Class Proportions

Second, we consider the case where the data class proportions are unknown or unclear. This case is relatively rare but does occur. For example, Stewart, Bosch, Chen, et al. (2017) detected students' wandering minds from facial expressions, with ground truth labels derived from self-reports. However, students only self-reported positive instances, so the negative instances were drawn from periods of time well before or after the self-reports. The actual rate of occurrence of positive instances is unclear in this situation, so the authors estimated the data class proportions from related literature.

A similar situation arises when a student model is not derived from data at all but must be applied to unseen data. For example, a simple model might be created where a computer programming student is predicted as frustrated if they experience three compiler errors in one minute. The data class proportions and predicted class proportions are both unknown for this model until some data is collected, at which point the predicted class proportions could be determined from the model, but the data class proportions would remain unclear until frustration labels were obtained from another source (e.g., expert annotation). In this case the chance-level precision, recall, and F_1 formulations are based on approximations of data and prediction class proportions. Thus, it is important to note when testing and reporting models that chance levels for these metrics will also be approximate.

3.3. Evaluation of Metrics on Real and Simulated Student Models

We evaluate F_1, precision, recall, kappa, and accuracy on real and simulated student models to illustrate situations where these metrics provide different perspectives on the same models. For all models, we compare metrics as well as chance levels. First, we evaluate several student models reported in the literature. These models make discrete predictions of student state, such as affect and behaviour, and thus require discrete model metrics for evaluation. Confusion matrices for these models were obtained from publications where possible, or from the corresponding authors where confusion matrices were not reported.

3.3.1. Gaze-Based Mind Wandering Detection

Hutt et al. (2017) detected mind wandering, an attentional state in which thoughts drift away from the task at hand to task-unrelated topics — often without immediate awareness that attention has lapsed (Smallwood & Schooler, 2015). They collected self-reports of mind wandering from students with randomly triggered probes as the students interacted with an intelligent tutoring system in a classroom. Gaze was simultaneously recorded, and mind wandering was detected from their gaze patterns with a machine-learned model. There were 2,334 instances classified, with 23% reported as mind wandering. In this article, we consider only the best-performing model (according to F_1) from among models reported.

3.3.2. Face-Based Mind Wandering Detection

Stewart, Bosch, & D'Mello (2017) detected self-reported mind wandering from facial features in two laboratory studies. Students in one study watched a narrative film, while in the other study students read an instructional scientific text. One of their primary goals was to measure how well face-based mind wandering detectors generalized across task domains. Thus, they downsampled data from the two studies to have equal numbers of instances (1,100 each) and equal mind wandering rates (25%). In this article, we consider the model reported with the highest performance and generalization performance combined (F_1 within-domain + F_1 across-domain). This model was trained on data from the narrative film-watching task, but applied to both datasets, yielding two sets of predictions.

3.3.3. Face-Based Affect and Behaviour Detection

Bosch et al. (2016) detected several affective states (boredom, confusion, delight, engagement, frustration) and off-task behaviour in a classroom context, with machine-learned models and features derived from facial action units. Students played an educational game designed to teach fundamental physics concepts, while trained observers recorded their affective and behavioural states according to the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; Ocumpaugh, Baker, & Rodrigo, 2015). The base rates (data class proportions) of each affective state varied, from 2.7% for confusion to 74.7% for engagement. The number of instances available for classification varied slightly as well due to feature extraction differences between states, ranging from 1,003 to 1,385 instances.

3.3.4. Interaction-Based Detection of Gaming the System

Paquette, de Carvalho, and Baker (2014) identified 13 interaction patterns (sequences of actions taken by a student) that were predictive of students attempting to "game the system" in a computerized learning environment for teaching algebra. Gaming the system occurs when a student attempts to progress through a learning task by abusing the affordances of the learning environment (e.g., repeatedly pressing a hint button until the final answer is shown). A final model was constructed from the 13 interaction patterns and tested on a holdout set of 2,599 instances (6.8% gaming).

3.3.5. Interaction-Based Affect Detection

Botelho, Baker, and Heffernan (2017) detected confusion, concentration, boredom, and frustration from students' interaction-log files in a web-based learning platform, which was integrated in both classroom and homework contexts. Trained observers recorded students' affective states using BROMP. Data class proportions were highly prevalent — the least common affective state (frustration) occurred in just 3.6% of instances. Concentration was the most prevalent state at 80.4%. There were 7,663 BROMP observations, though after discarding observations with low observer confidence and instances where there was no interaction data, final models were tested on 2,633 instances.

3.3.6. Simulated Student Models

In addition to the models mentioned above, we constructed simulated models so that the effects of data class proportions and predicted class proportions could be systematically varied and compared. We constructed three

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

types of simulated models. First, we fixed the data positive class proportion at .202 (mean of real student models examined) and varied the predicted positive class proportion to investigate the effect that prediction imbalance has on chance levels for precision, recall, and F_1. Second, we varied both data class proportions and predicted class proportions to show the interaction of the two types of class proportions on chance-level F_1. Finally, we also created simulated models with best-possible performance for various predicted class proportions with data class proportions fixed, to quantify the reduction in maximum performance due to predicted class proportions.

Table 1 illustrates an example confusion matrix for the type of simulated model created to measure best-possible performance in the presence of imbalanced predicted class proportions. In this example, the data positive class proportion was 50% (500 positive instances out of 1,000 total), the predicted positive class proportion was 20% (200 positive predictions), and correctness was as high as possible given these imbalances. Kappa in this example was .400 and F_1 was .571.

Table 1. Example Simulated Model and Dataset

		Predicted	
		Positive	Negative
Actual	Positive	200	300
	Negative	0	500

4. Results

We first present results calculated from previously published student models, then systematically explore chance levels with simulated models.

4.1. Example Student Models Built with Real Data

Table 2 contains confusion matrices for each student model, metrics computed from the matrices, and chance levels for precision, recall, and F_1. Several relevant trends are apparent in these models. First, the positive class is generally over-predicted in these models, which can be seen by comparing chance precision (data positive class proportion) to chance recall (predicted positive class proportion). In the most extreme example, the face-based boredom model predicted 37% of instances as the positive class while the rate of boredom in the data was just 4%. In fact, the only model where the positive class was not equal or over-predicted was the face-based engagement detection model, where the positive class was not the minority class. Mean data positive class proportion in these models was 20.2%, while mean predicted positive class proportion was 36.4%. This is a common pattern for student models with data imbalance and difficult prediction tasks, in that the positive class (the class of interest) is over-predicted to minimize false negatives or improve performance.

Second, the variance in data and prediction imbalances led to a large range of F_1 chance values, from .04 to .82. This illustrates how important it is to consider F_1 chance values carefully when reporting model performance or selecting a best model. An F_1 of .73 for the face-based engagement detection model might appear outstanding at first glance, but observing that a random number generator with the same prediction imbalance would result in F_1 = .66 (chance level) indicates that this result is more modest. Conversely, the interaction-based “gaming the system” detection model has F_1 of only .39, which may appear relatively small, but chance level F_1 is just .09. In fact, this model has the largest improvement in F_1 versus chance level in terms of absolute difference.

Third, the difference between accuracy and a simple majority baseline (predicting all instances as the majority class) further illustrates the importance of considering prediction imbalance, not only data imbalance. Model accuracies were actually below the majority baseline in every case except the interaction-based concentration detector. The majority baseline accuracy was overly conservative for these models, particularly where some over-prediction of the positive class was acceptable or even desirable given the difficulty of a prediction task. Kappa, on the other hand, subtracts an accuracy baseline that is relative to both prediction and data imbalances.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Table 2. Metrics from Example Student Models Applied to Real Student Data

Model	TP	FN	FP	TN	Precision	Chance Precision	Recall	Chance Recall	F_1	Chance F_1	Kappa	Accuracy	Majority Accuracy
Gaze													
Mind wandering	.161	.069	.431	.339	.27	.23	.70	.59	.39	.33	.09	.50	.77
Face													
Mind wandering (within domain)	.194	.056	.445	.305	.30	.25	.78	.64	.44	.36	.12	.50	.75
Mind wandering (cross-domain)	.190	.060	.493	.257	.28	.25	.76	.68	.41	.37	.07	.45	.75
Boredom	.024	.017	.347	.613	.06	.04	.58	.37	.12	.07	.05	.64	.96
Confusion	.011	.016	.245	.729	.04	.03	.42	.26	.08	.05	.03	.74	.97
Delight	.021	.009	.161	.809	.11	.03	.69	.18	.20	.05	.15	.83	.97
Engagement	.489	.258	.099	.154	.83	.75	.66	.59	.73	.66	.22	.64	.75
Frustration	.084	.059	.320	.537	.21	.14	.59	.40	.31	.21	.12	.62	.86
Off task	.029	.016	.171	.783	.15	.05	.65	.20	.24	.07	.18	.81	.95
Interaction logs													
Gaming the system	.036	.032	.081	.851	.31	.07	.53	.12	.39	.09	.33	.89	.93
Confusion	.006	.031	.044	.919	.12	.04	.16	.05	.13	.04	.09	.93	.96
Concentration	.730	.074	.115	.081	.86	.80	.91	.85	.89	.82	.35	.81	.80
Boredom	.047	.076	.079	.797	.37	.12	.39	.13	.38	.13	.29	.85	.88
Frustration	.007	.029	.035	.930	.17	.04	.19	.04	.18	.04	.15	.94	.96

Note: TP, FN, FP, and TN refer to proportion of all instances predicted as true positives, false negatives, false positives, and true negatives respectively.

Finally, these models exhibited some complexity that was not captured well by kappa or F_1 alone. The two models with most similar kappa values were the gaze-based mind wandering detection model (kappa = .09) and the interaction-based confusion detection model (kappa = .09). Conversely, F_1 for the mind wandering model was .39, versus .04 for the confusion model. Some of the difference in F_1 is explained by data imbalance, which was 4% confusion versus 23% mind wandering. However, the mind wandering model predicts 59% of instances as the positive class (mind wandering), while the confusion model over-predicted much less, with just 5% of instances predicted as the positive class. These differences between the two models were apparent in F_1, and especially in precision and recall, but might have been missed in a comparison focused on kappa. On the other hand, the two models with closest F_1 scores were the gaze-based mind wandering model (F_1 = .39) and the interaction-based gaming the system model (F_1 = .39). However, the gaming model was much better than chance (.39 vs. .09) while the mind wandering model was only modestly above chance (.39 vs. .33). Comparison of kappa (.09 vs. .33) or chance level F_1 of these two models makes the difference apparent.

4.1.1. Simulated Student Modelling Tasks

We constructed simulated models to systematically evaluate and illustrate key trends that were apparent in the real student models. First, we studied the effect of prediction imbalance on chance-level random models. Figure 2 shows how chance recall, chance precision, and chance F_1 vary as prediction imbalance increases when data imbalance is held constant at 20.2%, which was the mean data imbalance of the real student models discussed above.[2] Prediction imbalance influences recall chance level linearly, because in a random student model there is a linear relationship between how many positive predictions are made and how many instances of the positive class are correctly identified.

Importantly, the influence of predicted class proportions on chance-level F_1 is the same as the influence of data class proportions. Figure 3 demonstrates the symmetrical effect of imbalances on chance-level F_1 in more detail, varying both predicted and data class proportions in a model making random predictions. Chance-level is symmetrical along the diagonal. Figure 3 demonstrates another pattern, which is that the change in F_1 chance level due to predicted class proportion is most notable when the data positive class proportion is also high (viz. the positive class is also the majority class). Similarly, data class proportions affect chance-level F_1 most dramatically when the predicted positive class proportion is high.

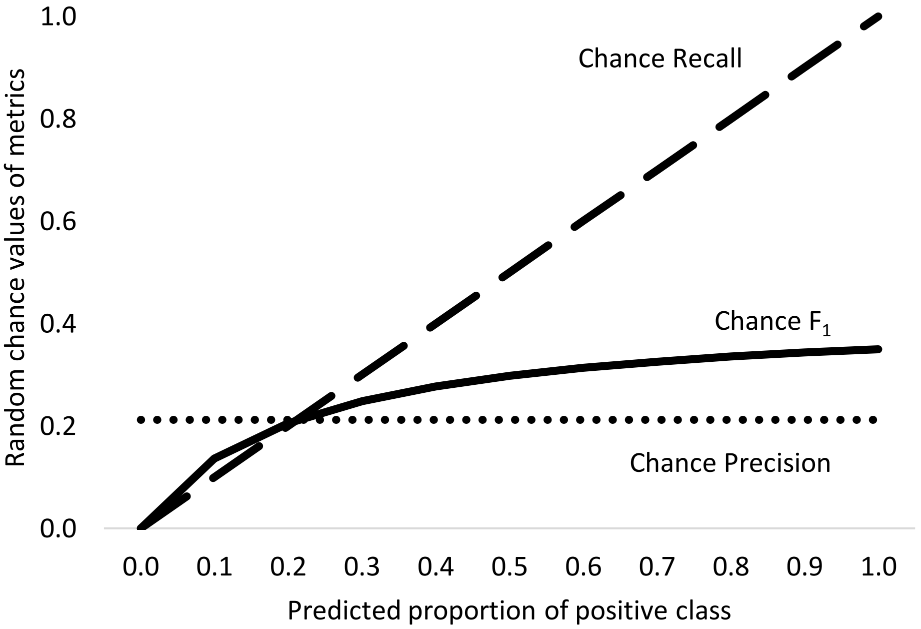

Figure 2. Effect of prediction imbalance on chance levels, with data imbalance fixed at 20.2%.

[2] 20.2% was chosen as a representation of common data class proportions, but chance levels vary due to predicted class proportions regardless of data class proportions. Even for a model with unequal data class proportions, it is possible to increase the recall and F_1 chance levels of a random model by increasing the predicted positive class proportion.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

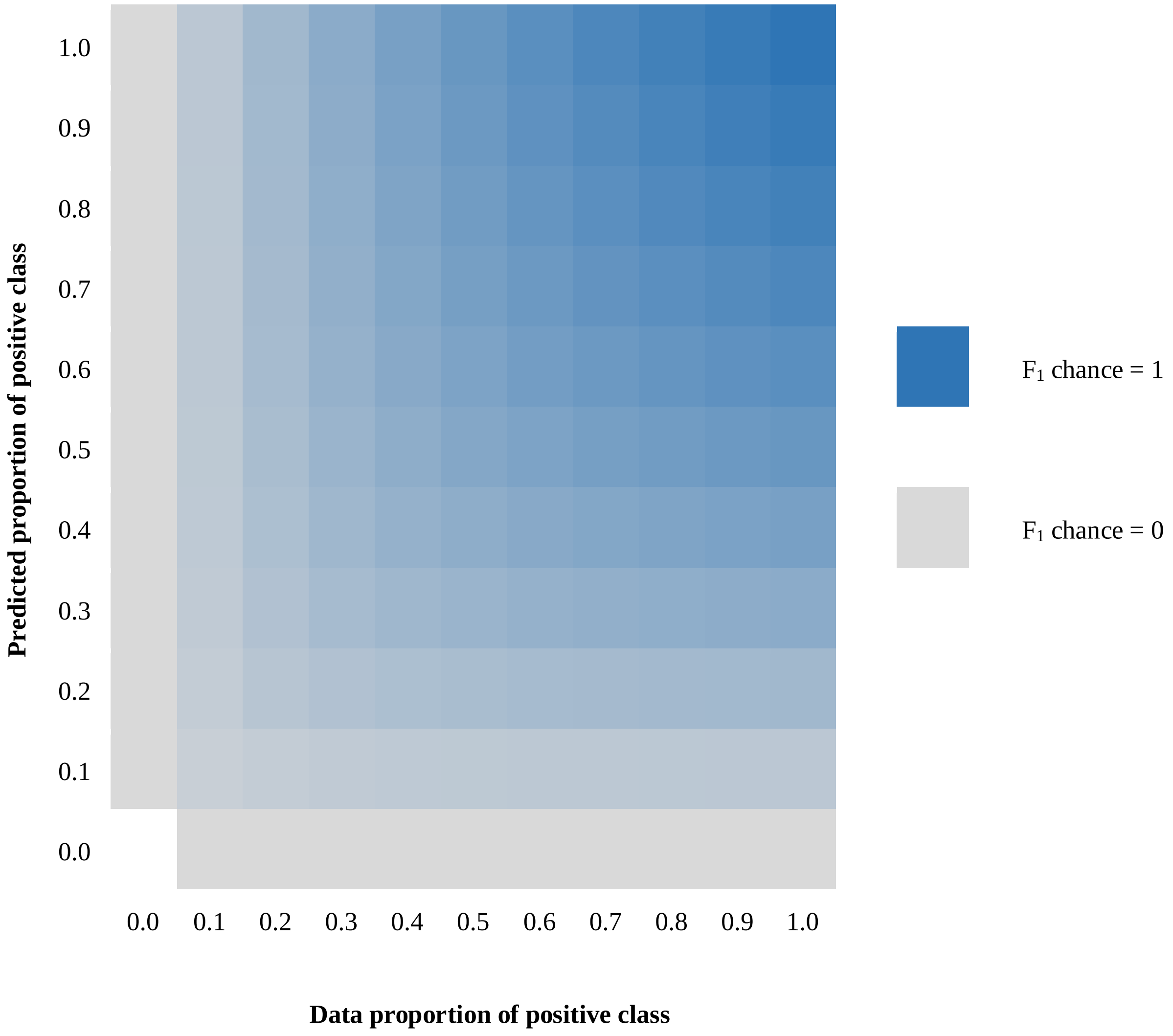

Figure 3. Chance level of F_1 while varying both data imbalance and prediction imbalance. Chance level is undefined when there are no instances or predictions of the positive class.

We also systematically explored the influence of predicted class proportions on peak model performance by simulating best-possible student models that always achieve the best possible performance given the predicted class proportions, shown in Figure 4 and Figure 5. These models were generated by calculating the maximum possible true positive and true negative rates of a model that predicts the positive class at a certain rate. For example, over-predicting models always had perfect recall but imperfect precision (due to false positives), while under-predicting models always had perfect precision but imperfect recall (due to false negatives).

First, Figure 4 illustrates a situation where the data class proportions are highly imbalanced such that positive class is the minority. Data positive class proportion was fixed at 20.2%, which was the mean proportion in the real student models discussed above. F_1 and kappa were maximized when predicted positive class proportion was also 20.2%, because the best possible model makes no false positives due to over-prediction or false negatives due to under-prediction. However, the best possible performance was notably diminished when predicted class proportions were imbalanced relative to data, as is often the case with student models. Furthermore, the best possible kappa and F_1 scores were similar near the point at which prediction and data class proportions matched but diverged as the predicted positive class proportion increased.

Conversely, when data class proportions were highly imbalanced such that the positive class was the majority (90% in this example), different patterns emerged (Figure 5). Most notably, best possible F1 differed dramatically versus best possible kappa across most predicted class proportions, while with a minority positive class (Figure 4) the difference was less drastic. Additionally, the difference between chance-level F1 and best possible F1 was less when the positive class was the majority.

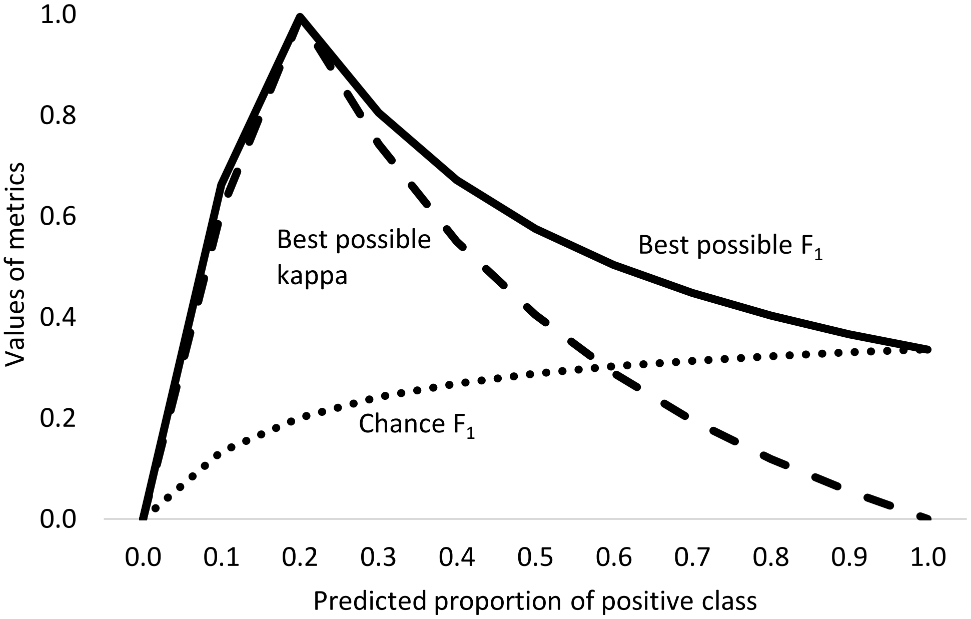

Figure 4. Performance of the best possible model with data imbalance fixed at 20.2% and varied prediction imbalance.

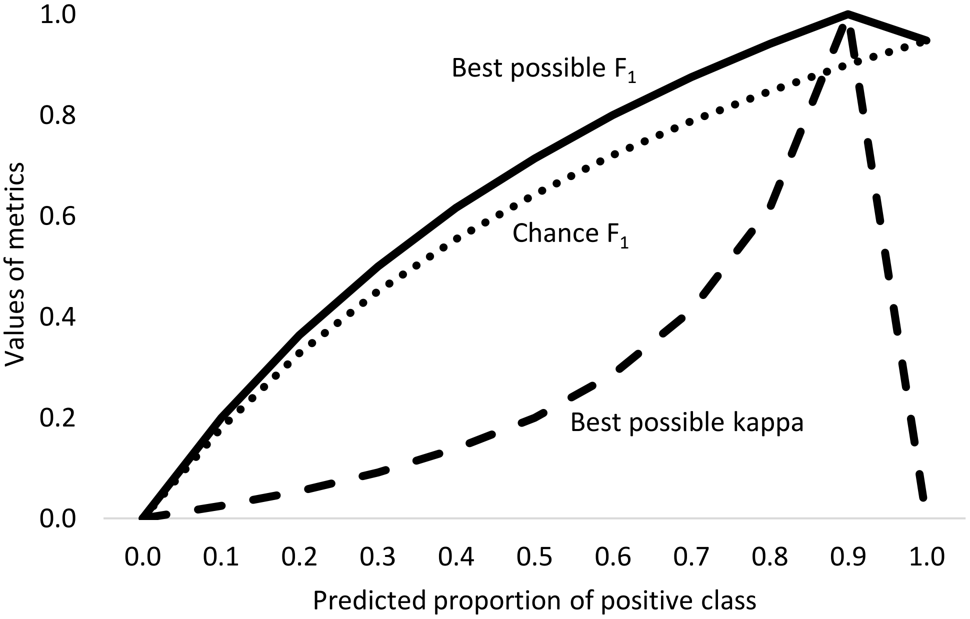

Figure 5. Performance of the best possible model with data imbalance fixed at 90% and varied prediction imbalance.

5. General Discussion

We were interested in systematically evaluating the biases of key performance metrics (precision, recall, F_1, kappa, and accuracy) for student models with discrete predictions and outcomes. To explore this topic, we provided mathematical

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

formulations for random chance-level F_1 baseline and evaluated F_1 and other metrics on real and simulated models with differing data class proportions and predicted class proportions.

5.1. Main Findings

From previous research (Jeni et al., 2013), we expected F_1 would be influenced by imbalanced data class proportions (skew). We found that even in an imbalanced dataset, the maximum value of F_1 is not attenuated if the prediction imbalance exactly matches data imbalance, but in a more realistic scenario with over-prediction of the positive class, it is indeed (Figure 4). We also found that the random chance level of F_1 is influenced by prediction. Specifically, a model making random predictions can increase its F_1 score by predicting a higher proportion of the positive class. For instance, in the Figure 2 example with data positive class proportion = .202, a random model with matching predicted class proportions will have F_1 = .202. However, the random model F_1 can be increased to .350 simply by over-predicting the positive class.

We also noted, however, that over-prediction is common in examples of published discrete student models (Table 1). Indeed, over-prediction may often be desirable to minimize false negatives, since some infrequent student states have important relationships with learning (Baker et al., 2010; Beck & Rodrigo, 2014; Cocea, Hershkovitz, & Baker, 2009; Pardos et al., 2013). Therefore, it is important to acknowledge that F_1 (and recall) may be positively skewed by over-prediction, and to compare to the appropriate chance level. Our analyses also showed that accuracy often falls below the majority baseline due to over-prediction of the minority class (Table 2), and thus might not be the most appropriate metric for student models.

Simulated student models demonstrated the effect that predicted class proportions have on performance. For common values of data class proportions, the best possible F_1 was reduced from 1 to as low as .336 by over-prediction, while kappa was reduced from 1 to as low as 0 (Figure 4). It is no surprise that over-prediction lowers maximum performance, but it is important to note that both kappa and F_1 were reduced to zero by total under-prediction of the positive class, while F_1 was not reduced to zero (but kappa was) by over-prediction. As such, interpretation of F_1 is particularly dependent on whether the model is over- or under- predicting. Furthermore, when the positive class was the majority, we noted that the best possible F_1 differed little from chance-level F_1 (Figure 5) and both were higher than the best possible kappa for almost all levels of predicted class proportions. Thus, if measuring model performance with F_1 in a dataset with a large positive class majority (e.g., 90% in Figure 5), it should be noted that F_1 will not greatly exceed chance level even in a relatively accurate model.

5.2. Implications

Some recommendations for examination of metrics emerge from these findings. First, a single metric by itself is likely to either hide important differences between models, as illustrated in Table 2 where models with similar kappa could result in very different F_1, and vice versa. Thus, it may be helpful to examine both kappa and F_1 or to report chance level F_1 along with F_1. Furthermore, interpretation of kappa is especially difficult as it captures positive and negative classes in a model equally, while F_1 frames performance in terms of the positive class alone. However, both F_1 and kappa are cumbersome to interpret as the definitions are complex (equations 1 and 4). Precision (equation 2) and recall (equation 3), on the other hand, are simple ratios of intuitive values that can lend interpretability to a model.

Examining and reporting multiple metrics for student models is thus likely to be illuminating, but is not a panacea for model evaluation and reporting issues. Reporting multiple metrics does not avoid the problem of reporting chance level — rather, multiple chance levels should also be reported (e.g., one each for precision, recall, and F_1 if all three are reported). Furthermore, in certain situations it may be necessary to evaluate models based on a single number. This is the case when one unique model must be selected out of many. For example, single-metric comparisons can also occur during the training of a single model when tuning hyperparameters with nested cross-validation, as is the case with forward feature selection (Guyon & Elisseeff, 2003). In such situations, it is not possible to compare models across multiple metrics, because one model may be better according to one metric and worse according to a different metric. It is therefore sometimes necessary to rank models by a single metric chosen to favour model goals such as minimizing false positives or false negatives. One could also combine metrics via an averaging function (e.g., as F_1 is to precision and recall) in an attempt to select models avoiding the pitfalls of individual metrics. However, the chance level of this averaged metric will have to be calculated and considered. For example, combining F_1 and AUC by adding them effectively creates a new single metric with its own chance level (F_1 chance + AUC chance) that needs to be accounted for.

The choice of metric for ranking, selecting, and optimizing models influences the results. Thus, choosing an appropriate metric is crucial. We found that the maximum possible kappa was highest when predicted and data class proportions were equal, yet predicted class proportions do not affect kappa. Thus, kappa is likely a suitable metric for model selection in situations where matching predicted class proportions to data is desirable. Optimizing a model based on recall will likely lead to excessive over-prediction, since recall is maximized when a model predicts everything as the positive class. Hence, F_1 may

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

be a better choice when false negatives should be minimized, since it balances recall with precision. Precision penalizes false positives alone, and thus is an appropriate choice when false positives should be minimized. For example, a computerized learning environment might administer a relatively large intervention (such as giving the student a 5-minute break) if the student is bored, but only if the student model has a small chance of predicting false positives. Conversely, a model that over-predicts would be more suitable for "fail-soft" interventions such as discretely alerting a teacher to observe the student and assess the need for further action. Furthermore, for models with underlying continuous predictions (e.g., logistic regression), selecting a decision threshold (e.g., make a positive prediction for values > .5 and vice versa) is a similar problem of optimization. Selecting the decision threshold to maximize F_1, for example, may result in a model that over-predicts the positive class, since F_1 can be increased by chance through over-prediction.

We found that F_1 chance levels were influenced by predicted and data class proportions which should thus be considered in parallel with F_1 when examining model performance. We found that normalizing F_1 relative to chance level in the same way as kappa (i.e., subtracting chance and dividing by $1 - chance\ F_1$) resulted in a metric equivalent to kappa. There are, however, other choices of F_1 normalization that could be considered in addition to the method analogous to kappa. Normalizing with the $1 - chance\ F_1$ denominator will scale the metric value in the range [–1, 1] only in the event of no over- or under- prediction. However, as seen in Figure 4, the maximum possible value of F_1 can be significantly decreased by over-prediction. If some over-prediction is acceptable to avoid false negatives, then such normalization may be too conservative. F_1 could instead be normalized relative to the maximum value possible for the level of over-prediction present, i.e., $max\ possible\ F_1 - chance\ F_1$. This normalization would be interpreted as the proportion of the distance between chance level and maximum possible F_1, *for the model's level of over-prediction*, that was covered by the model's F_1. More work is needed in the future to discover where this method and other normalization methods are effective, or not conservative enough, and how they compare to other metrics.

Lastly, we calculated appropriate chance levels for precision, recall, and F_1 in examples of previously published work where confusion matrices were reported or could be inferred from the results (Table 2). This suggests that a large-scale systematic review of student modelling literature is possible and could uncover trends that would further inform model evaluation practices. For instance, do models optimized for F_1 tend to over-predict the positive class more than models optimized for kappa? Furthermore, it also suggests that reporting confusion matrices is a good practice, as it allows post-hoc calculation of discrete model metrics (and, in this case, chance levels) not considered by researchers before publication.

5.3. Limitations and Future Work

The evaluation of metrics for discrete models in this article was not without limitations. We identified several key limitations that also present opportunities for future work. First, while the discussed metrics are perhaps the most commonly reported in discrete student modelling research (Gardner & Brooks, 2017), there are other metrics worth investigating. For example, it is unclear how suitable Matthew's Correlation Coefficient (MCC) might be for evaluating student model performance in cases of data and prediction imbalances. Future work should replicate this investigation with MCC and other metrics that may be suitable for evaluating discrete student models, such as Bangdiwala's B. Additionally, measures such as precision, recall, and F_1 that measure class-specific performance (almost always the positive class) can also be averaged across classes. Similarly, Pelánek (2017) notes that metrics are commonly averaged not only across classes but also across students and other levels of the data. Further research will be needed to understand the biases and chance levels of discrete model metrics in such use cases.

Second, while the results discussed here provided some guidelines about when certain metrics would be more applicable than others for student modelling goals, it would be best to measure empirically the effect that choice of metric has on model ranking and selection. A larger number of published student models and simulated models should be created to explore precisely the effect of metric choice on the features selected in forward (and backward) feature selection, as well as fully trained models. For example, it is unclear whether choice of metric would have a significant influence on the number of features selected, or on the type of features selected.

Third, we did not consider methods for calculating statistical significance of metrics relative to chance levels. For some metrics (e.g., kappa) the significance is well-defined. For others, such as F_1, repeated randomization is often employed to test significance (Yeh, 2000). Statistical significance testing may be desirable in student modelling research, and ease of testing could perhaps influence choice of metrics. Furthermore, the relationship between F_1 and kappa shown in this article suggests the existence of a closed-form statistical significance test for F_1 that should be explored in future work.

Finally, we considered metrics only for student models that produce discrete outcome predictions, but there are many student models that produce continuous predictions. We focused on discrete model metrics because of their prevalence in student modelling literature, and because related research had focused primarily on continuous metrics (e.g., Jeni et al., 2013;

Pelánek, 2015). However, there are still open questions for continuous metrics, such as the best way to choose a decision threshold for making discrete predictions from continuous predictions. In one example, Stewart, Bosch, & D'Mello (2017) optimized the threshold for F_1, but optimizing for another metric might yield a different model. In future work, we will investigate choice of metric for threshold fitting more systematically as well, to inform best practices for student modelling.

5.4. Concluding Remarks

A variety of metrics are reported in research describing discrete student model evaluations, but their relationships with each other and with random chance level performance are not always made clear. We evaluated metrics on various published student models to uncover situations where metrics agree and disagree about student model performance, and what the implications are for selecting and reporting metrics. We found that predicted class proportions were especially influential on the values of recall and F_1 for student models with over-prediction of the positive class, and established random chance levels for these metrics that account for prediction imbalance. Our findings provide some guidance on best practices (e.g., reporting chance level F_1) and suggest fruitful opportunities for future research on the influence of metrics in all aspects of student model engineering. Eventually, a full evaluation of these metrics will lead to student models that are better suited for their specific purpose, and in turn improve instruction and understanding of the learning process.

Acknowledgements

We would also like to thank the researchers who published the student models that we evaluated in this paper, for providing clarifications and additional details about models where needed.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Please insert relevant information here

Funding

Nigel Bosch was supported by postdoctoral research funding provided by the National Center for Supercomputing Applications.

References

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708. http://dx.doi.org/10.1016/j.chb.2005.12.009

Baker, R. S., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In B. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (ITS 2008), 23–27 June 2008, Montreal, PQ, Canada (pp. 406–415). Springer. http://dx.doi.org/10.1007/978-3-540-69132-7_44

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students "game the system." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '04), 24–29 April 2004, Vienna, Austria (pp. 383–390). New York: ACM. http://dx.doi.org/10.1145/985692.985741

Baker, R. S., D'Mello, S. K., Rodrigo, M. 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. http://dx.doi.org/10.1016/j.ijhcs.2009.12.003

Beck, J., & Rodrigo, M. M. T. (2014). Understanding wheel spinning in the context of affective factors. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (ITS 2014), 5–9 June 2014, Honolulu, HI, USA (pp. 162–167). New York: Springer. http://dx.doi.org/10.1007/978-3-319-07221-0_20

Bixler, R., & D'Mello, S. K. (2015). Automatic gaze-based detection of mind wandering with metacognitive awareness. In F. Ricci, K. Bontcheva, O. Conlan, & S. Lawless (Eds.), *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization* (UMAP 2015) 29 June–3 July 2015, Dublin, Ireland (pp. 31–43). Springer. http://dx.doi.org/10.1007/978-3-319-20267-9_3

Bosch, N., Crues, R. W., Henricks, G. M., Perry, M., Angrave, L., Shaik, N., Bhat, S., & Anderson, C. J. (2018). Modeling

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

key differences in underrepresented students' interactions with an online STEM course. *Proceedings of the Technology, Mind, and Society Conference* (APATech18), 5–7 April 2018, Washington, DC, USA. New York: ACM. http://dx.doi.org/10.1145/3183654.3183681

Bosch, N., D'Mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(2). http://dx.doi.org/10.1145/2946837

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. In E. André, R. S. Baker, X. Hu, M. M. T. Rodrigo, & B. du Boulay (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education* (AIED 2017), 28 June–1 July 2017, Wuhan, China (pp. 40–51). Springer. http://dx.doi.org/10.1007/978-3-319-61425-0_4

Bower, G. H. (1992). How might emotions affect learning? *The Handbook of Emotion and Memory: Research and Theory*, 3–31. Hillsdale, NJ: Lawrence Erlbaum.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. http://dx.doi.org/10.1109/T-AFFC.2010.1

Cetintas, S., Si, L., Xin, Y. P. P., & Hord, C. (2010). Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. *IEEE Transactions on Learning Technologies*, *3*(3), 228–236. http://dx.doi.org/10.1109/TLT.2009.44

Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, *7*(3), 246–259. http://dx.doi.org/10.1109/TLT.2013.2296520

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729. http://dx.doi.org/10.1016/j.eswa.2013.02.007

Cocea, M., Hershkovitz, A., & Baker, R. S. (2009). The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (AIED '09) 6–10 July 2009, Brighton, UK (pp. 507–514). Amsterdam, Netherlands: IOS Press.

Cohen, W. W. (1995). Fast effective rule induction. In A. Prieditis & S. Russell (Eds.), *Proceedings of the 12th International Conference on Machine Learning* (ML95), 9–12 July 1995, Tahoe City, California (pp. 115–123). San Francisco, CA: Morgan Kaufmann. http://dx.doi.org/10.1016/B978-1-55860-377-6.50023-2

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1–2), 9–38. http://dx.doi.org/10.1007/s11257-011-9106-8

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*(Mar), 1289–1305.

Gardner, J., & Brooks, C. (2017). Student success prediction in MOOCs. *ArXiv:1711.06349 [Cs, Stat]*. http://arxiv.org/abs/1711.06349

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*(1), 36–53. http://dx.doi.org/10.1016/j.compedu.2015.09.005

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-Measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298. http://dx.doi.org/10.1197/jamia.M1733

Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J., & D'Mello, S. K. (2017). Out of the fr-"eye"-ing pan: Towards gaze-based models of attention during learning with technology in the classroom. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (UMAP 2017), 9–12 July 2017, Bratislava, Slovakia (pp. 94–103). New York: ACM. http://dx.doi.org/10.1145/3079628.3079669

Jeni, L. A., Cohn, J. F., & De la Torre, F. (2013). Facing imbalanced data: Recommendations for the use of performance metrics. *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (ACII '13), 2–5 September 2013, Geneva, Switzerland (pp. 245–251). IEEE Computer Society. http://dx.doi.org/10.1109/ACII.2013.47

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. *Proceedings of the IEEE International Conference on Advanced Learning Technologies* (ICALT 2001), 6–8 August 2001, Madison, WI, USA (pp. 43–46). IEEE Computer Society.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

http://dx.doi.org/10.1109/ICALT.2001.943850

Lawvere, F. W. (1973). Metric spaces, generalized logic, and closed categories. *Rendiconti Del Seminario Matématico e Fisico Di Milano*, *43*, 135–166.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. http://dx.doi.org/10.1111/j.1466-8238.2007.00358.x

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196–204. http://psycnet.apa.org/doi/10.1037/a0014104

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. *Proceedings of the 9th International Conference on Spoken Language Processing* (INTERSPEECH 2006 — ICSLP), 17–21 September 2006, Pittsburgh, PA, USA (pp. 809–812). International Speech Communication Association.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1–32. http://dx.doi.org/10.1016/0010-0285(87)90002-8

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

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64.

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (CogSci 2014), 23–26 July 2014, Quebec City, Canada (pp. 1126–1131). Austin, TX: Cognitive Science Society.

Pardos, Z. A., Baker, R. S., San Pedro, M. O. C. Z., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 117–124). New York: ACM. http://dx.doi.org/10.1145/2460296.2460320

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105. http://dx.doi.org/10.1207/S15326985EP3702_4

Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining*, *7*(2), 1–19.

Pelánek, R. (2017). Measuring predictive performance of user models: The details matter. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (UMAP 2017), 9–12 July 2017, Bratislava, Slovakia (pp. 197–201). New York: ACM. http://dx.doi.org/10.1145/3099023.3099042

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, Informedness, Markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Raca, M., Kidzinski, L., & Dillenbourg, P. (2015). Translating head motion into attention: Towards processing of student's body-language. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 320–326). International Educational Data Mining Society.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). Forecasting student achievement in MOOCs with natural language processing. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 383–387). New York: ACM. http://dx.doi.org/10.1145/2883851.2883932

Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G., & Gurcan, M. N. (2013). Mitosis detection in breast cancer histological images: An ICPR 2012 contest. *Journal of Pathology Informatics*, *4*. http://dx.doi.org/10.4103/2153-3539.112693

Smallwood, J., Fishman, D. J., & Schooler, J. W. (2007). Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*, *14*(2), 230–236. http://dx.doi.org/10.3758/BF03194057

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, *66*(1), 487–518. http://dx.doi.org/10.1146/annurev-psych-010814-015331

Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, *3*(2), 211–223. http://dx.doi.org/10.1109/T-AFFC.2011.37

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Stewart, A., Bosch, N., Chen, H., Donnelly, P. J., & D'Mello, S. K. (2017). Face forward: Detecting mind wandering from video during narrative film comprehension. In E. André, R. S. Baker, X. Hu, M. M. T. Rodrigo, & B. du Boulay (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education* (AIED 2017), 28 June–1 July 2017, Wuhan, China (pp. 359–370). Springer. http://dx.doi.org/10.1007/978-3-319-61425-0_30

Stewart, A., Bosch, N., & D'Mello, S. K. (2017). Generalizability of face-based mind wandering detection across task contexts. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 88–95). International Educational Data Mining Society.

Trigwell, K., Ellis, R. A., & Han, F. (2012). Relations between students' approaches to learning, experienced emotions and outcomes of learning. *Studies in Higher Education*, *37*(7), 811–824. http://dx.doi.org/10.1080/03075079.2010.549220

Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., & Scherer, K. (2012). Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(4), 966–979. http://dx.doi.org/10.1109/TSMCB.2012.2200675

Walonoski, J. A., & Heffernan, N. T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In M. Ikeda, K. Ashlay, & T.-W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (ITS 2006), 26–30 June 2006, Jhongli, Taiwan (pp. 382–391). Springer. http://dx.doi.org/10.1007/11774303

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. *Proceedings of the 18th Conference on Computational Linguistics* (COLING '00), 31 July–4 August 2000, Saarbrücken, Germany (Vol. 2, pp. 947–953). Stroudsburg, PA: Association for Computational Linguistics. http://dx.doi.org/10.3115/992730.992783

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)