Population validity for Educational Data Mining models: A case study in affect detection

Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan

Jaclyn Ocumpaugh (PhD, Michigan State University) is a Research Associate at Teachers College, Columbia University, where her research focuses on making educational technologies culturally accessible. Ryan Baker (PhD, Carnegie Mellon University) is an Associate Professor of Cognitive Studies in Education at Teachers College, Columbia University, the founding President of the International Educational Data Mining Society, and the Associate Editor of the Journal of Educational Data Mining. Sujith Gowda is a Research Programmer at Arizona State University. His research focuses on building automated machine learned models to detect robust learning, affective states and dis-engaged behaviors of students using educational software systems. Neil Heffernan (PhD, Carnegie Mellon University) is an Associate Professor of Computer Science and co-director of the PhD program in Learning Sciences and Technologies at Worcester Polytechnic Institute, where he developed the ASSISTments online learning system designed to improve mathematics education. He freely distributes ASSISTments, providing a tool for improving instruction and a platform for educational research. Cristina Heffernan has taught mathematics in both the United States and abroad. As co-founder of ASSISTments, she has been instrumental in nurturing the system to be a tool for teachers. Email: jo2424@columbia.edu; jocumpaugh@wpi.edu

Abstract:
ICT-enhanced research methods such as educational data mining (EDM) have allowed researchers to effectively model a broad range of constructs pertaining to the student, moving from traditional assessments of knowledge to assessment of engagement, meta-cognition, strategy, and affect. The automated detection of these constructs allows EDM researchers to develop intervention strategies that can be implemented either by the software or the teacher. It also allows for secondary analyses of the construct, where the detectors are applied to a data set that is much larger than one that could be analyzed by more traditional methods. However, in many cases, the data used to develop EDM models is collected from students who may not be representative of the broader populations who are likely to use ICT. In order to use EDM models (automated detectors) with new populations, their generalizability must be verified. In this study, we examine whether detectors of affect remain valid when applied to new populations. Models of four educationally relevant affective states were constructed based on data from urban, suburban, and rural students using ASSISTments software for middle school mathematics in the Northeastern United States. We find that affect detectors trained on a population drawn primarily from one demographic grouping do not generalize to populations drawn primarily from the other demographic groupings, even though those populations might be considered part of the same national or regional culture. Models constructed using data from all three sub-populations are more applicable to students in those populations than those trained on a single group, but still do not achieve ideal population validity—the ability to generalize across all sub-groups. In particular, models generalize better across urban and suburban students than rural students. These findings have important implications for data collection efforts, validation techniques, and the design of interventions that are intended to be applied at scale.

Practitioner Notes:

What is known about this topic:

- Educational data mining (EDM) techniques have enabled assessment ("detection") of a range of constructs, including ill-defined constructs such as disengagement and affect.
- Affect detection often leverages physiological sensors, but recent research has been able to detect affect solely from log files.
- EDM detection of any construct is often difficult to generalize from one platform to another or from one population to the next. (That is, few obtain population validity.)

What this paper adds:

- This paper expands knowledge of how far affect detection can be taken out of the original contexts in which models were developed.
- We find that diverse populations must be used to develop detectors, in order to guarantee that detectors will work with diverse populations.
- However, for two of the three populations in this study, population-general models can detect affective states as accurately as models that are trained only on a single population.

Implications for practice and/or policy:

- Educational Data Mining (EDM) techniques allow affect detection *in situ*, increasing the opportunities for introducing more sophisticated assessment, for informing teachers, and for automated intervention.
- Detection of any construct via EDM techniques allows for further research using discovery with models, a technique that allows for the rapid scaling of research.
- The population validity demonstrated in this study increases the likelihood that these detectors will scale to broader populations.

Introduction

In recent years, education researchers have become increasingly interested in methods used for harnessing data from information and communication technologies (ICT; Markauskaite, Kennan, Richardson, Aditomo, and Hellmers, 2012). Calls to develop new methods that exploit the fine-grained log files produced by ICTs have been an important theme in education research since the 1980s (see Anderson, 1983). Since then, a range of research communities have shown the benefit of ICT-enhanced research methods for education research, analyzing everything from teacher behavior (Smeets and Mooij, 2001) to student engagement (Rappa, Yip, and Baey 2009) to online collaborative learning (Reimann, Thompson, and Weinel, 2007).

One school of thought which has attempted to exploit data from ICTs is educational data mining (EDM; Romero & Ventura, 2007; Baker & Yacef, 2009). EDM is closely related to learning analytics (Ferguson, 2012). Each community leverages a combination of human judgment and automated analysis (Siemens and Baker, 2012) to understand learners and learning. Both EDM and learning analytics differ from traditional statistics in that they take an *a posteriori* approach to data. That is, EDM techniques automate the analyses of data in such a way that it is possible to discover structural patterns that might be missed if only a single theoretical construct or pre-selected hypothesis were being tested. As such, EDM techniques ameliorate the risks of confirmation bias that are inherent in traditional statistics (see Zhao and Luan, 2006). The differences between learning analytics and educational data mining are relatively minor and emerge in terms of philosophical orientation and emphasis. Specifically, learning analytics research often attempts to understand the learning system as a whole, in its full complexity, whereas EDM research often analyzes individual components or constructs in the processes surrounding learning and then attempts to understand the relationships and interactions

between these components or constructs (Siemens and Baker, 2012). The two also differ in the primary application of analyses, with learning analytics often having a greater focus on informing and empowering instructors and learners, whereas EDM often has a greater focus on basic discovery and automated adaptation.

Within the branch of EDM termed prediction modeling (Baker & Yacef, 2009), researchers distill large numbers of potential features of the data (the log files of the students' interactions with the ICT) and create a "ground truth" measure of some construct of interest (i.e. labels of the construct's presence or absence that are treated as essentially true despite their imperfections, such as human observations of student affect). Once those labels are applied to a small subset of the data, EDM researchers analyze the data with one or more algorithms to determine which features produce the most useful models of that construct. For example, human coders might have labeled whether or not individual students were experiencing boredom (and when) over a given observation period. EDM researchers would then use this ground truth label to determine which other data labels (features) in the software log files could be used to predict when a student is bored, producing a variety of models and then selecting those which appear most precise at distinguishing boredom. Those models can then be applied to much larger data sets so that human coders are no longer necessary.

Models developed from EDM prediction modeling methods, often called detectors, have been used to drive automated interventions and to provide timely information to instructors and students. For example, Baker et al. (2006) used prediction models to drive *in vivo* detection of students who were gaming the system, activating a cartoon pedagogical agent known as "Scooter the Tutor," who sought to put students back on track as they navigated the Cognitive Tutor software system. Developers of the same software system have also used prediction models of student knowledge to drive teacher reports (Johnson, 2005), allowing instructors to implement their own interventions from information predicted by EDM detectors without needing to understand how to create these models.

Increasingly, however, detectors are also being used to support discovery with models analyses, where a detector is then used to study the construct it models at a scale difficult to achieve with other educational research methods (see Hershkovitz, Baker, Gobert, Wixon, and Sao Pedro, 2013). That is, once a model is validated, we can use it to obtain a measure of the construct in databases of such an enormous scope that comprehensive human labeling would be infeasible, thus facilitating the discovery of patterns that would not be possible with more traditional methods of data collection and analysis. Such large-scale research, however, requires evidence on the validity of the models.

Previous educational research has addressed a wide range of variables that impact both internal and external validity (see Rupp, Gushta, Mislevy and Shaffer, 2010). Laboratory studies, which allow for highly controlled experiments, tend to favor methods that bias towards internal validity (confidence that the conclusions being drawn are warranted in the specific case being investigated). On the other hand, EDM techniques allow researchers greater opportunities to utilize data collected *in situ*, increasing opportunities for improving the external validity (the degree to which the findings can generalize to new contexts and populations, where generalization is defined as the same findings or models applying approximately as well for a new situation as for the original situation where the models or findings were developed).

As the field matures, EDM researchers have become more concerned with developing methods which improve validity. Recent research has sought to enhance theoretical construct validity (the degree to which the features of a model reflects the intended construct). For

example, Sao Pedro, Baker, and Gobert (2012) leveraged techniques from knowledge engineering research, using the judgments of domain experts to improve upon the automated feature selection process used to generate EDM models. This process helped to ensure that the features being used were qualitatively relevant and resulted in an improved model. Other researchers have leveraged the benefits of evidence-centered design methods, refining model design by improving upon the type and amount of data collected and then using the improved data to enhance EDM modeling (Rupp et al., 2012; Mislevy, Behrens, DiCerbo, and Levy 2012).

Other EDM researchers have begun to explore issues related to population validity, the ability of a model to generalize to new and distinct groups of students, particularly in respect to cross-cultural differences. For instance, San Pedro, Rodrigo, and Baker (2011) have investigated whether models of carelessness generalize between the USA and the Philippines (they do), and Soriano et al. (2012) have investigated international generalization of models of effective help-seeking, finding that models built on data from the USA and Philippines function effectively within the other country, but that models built using data from these countries do not work effectively in data from Costa Rica. In addition, research by Ogan et al. (2012) on the collaborative learning strategies used by Latin American students when using ICTs suggests that classic knowledge tracing algorithms developed in the USA may be less effective in Latin American cultural contexts. These results suggest that if EDM detectors are trained only on single populations, they are less likely to be effective when generalized to other groups. The process of building automated detectors using EDM can tolerate some noise in the ground truth (training) labels used to build models (see Baker et al., 2012), but it is very difficult to adjust if the data is systematically biased.

These broad intercultural cultural differences are clearly quite important if ICT-based education is going to be useful for the full diversity of learners worldwide, and we are hopeful that more researchers will answer Blanchard's (2012) call for EDM and related fields to move beyond the convenient populations from western, industrialized, wealthy countries (the so-called WEIRD countries) that are often used as research subjects. However, attention to inter-group differences that are finer-grained than national boundaries is also important (see Paez and Vergara, 2000) since there could be substantial intra-national differences between regions or even between groups within a region. Gender, race, and socioeconomic status are commonly investigated in sociological research, but categories like urbanicity may be particularly relevant to educational software usage patterns since previous research has shown that rural students may have less access to technology (Wilson, Wallin and Reiser, 2003; Chen and Liu, 2013). These and other differences may contribute to widening gaps in mathematics achievement in the United States that have also been shown to correlate with urbanicity (Wenglinsky, 1988; Graham and Provost, 2012).

In this paper, we present a case study of EDM research addressing these sub-national, inter-group differences among students in the Northeastern United States by looking at urban, suburban, and rural students use of the mathematics tutor ASSISTments. Specifically, we develop automated detectors that infer several educationally-relevant affective states from patterns in the students' log files. We then study whether these automated detectors can generalize from one population to another or whether an automated detector derived from all three populations works better. Our results suggest that perfect population validity may be difficult to achieve in some cases, and we discuss important methodological approaches that should be considered as a consequence.

Methods

Students in this study are drawn from middle schools in the Northeastern United States that represent three different populations: rural, suburban, and urban students. The first group, drawn from 2 schools in Maine, is comprised of a rural population of students from a fairly homogenous ethnic and socioeconomic background. These schools were more than 95% White and generally of lower socioeconomic status, with over half of students receiving a free or reduced-price lunch (a commonly-used indicator of poverty in the United States). The second group, drawn from three suburban schools in Massachusetts, is comprised predominantly of White and East-Asian students who are of mid-to-high socioeconomic status, with less than 20% of students receiving a free or reduced-price lunch. Finally, the third group, drawn from an urban setting in Massachusetts, is composed of largely of lower-income Latino/a students, mostly of Puerto Rican origin but speaking English as a first language, as well as a sizable minority of African-American students and students from the Balkans, with over half of students receiving a free or reduced-price lunch.

Dividing students in terms of urbanicity is useful in multiple fashions. First, it is common practice for research of this nature to be conducted using convenience samples, which typically eliminates either rural populations (due to distance) or disadvantaged urban populations (due to bureaucratic processes in urban schools as well as hazards to researchers conducting research in these regions). Second, urbanicity in the United States is often a proxy for ethnic, racial, and socioeconomic differences, and it is relatively common for research samples to exclude rural or urban groups, who often have more limited access to educational technology than their suburban peers. Third, it is a useful demographic for education research in its own right (Campbell, 1989; Hu, 2003).

Students in this study were observed using the ASSISTment system (Razzaq et al., 2005), a formative assessment system that provides hints and scaffolded instruction in mathematics. The ASSISTment system was developed at Worcester Polytechnic Institute and is typically used one day a week in class. It is increasingly also used for homework (Mendicino, Razzaq, and Heffernan, 2009).

[insert Figure 1 about here]

Within this paper, we discuss the development of automated detectors (prediction models) that infer student affect from log files of student interaction with the ASSISTment system. As discussed above, EDM detectors are developed using ground truth labels (e.g. human assessments) of the construct to be modeled are obtained. This data is used to train the model to infer the construct so that it can be detected even when human assessments are impossible or infeasible (either because it is too expensive to collect them or because future events are being predicted). In this paper, the ground truth labels of student affect are obtained using Quantitative Field Observations (QFOs) of educationally relevant affect categories. The QFOs were obtained using the Baker-Rodrigo Observation Method Protocol (BROMP; Ocumpaugh, Baker, & Rodrigo 2012), which achieves a higher level of inter-rater agreement (minimum Kappa of 0.6) for affect than traditional methods, such as Ekman's Facial Action Coding System (FACS; Sayette, Cohn, Wertz, Perrott, and Parrott, 2001) or video coding under comparable conditions (see D'Mello, Craig, Witherspoon, McDaniel, and Graesser 2008).

In this study, students' affective states were classified according to the following educationally-relevant categories: boredom (see Csikszentmihalyi, 1990; Miserandino, 1996),

confusion (see Craig, Graesser, Sullins and Gholson, 2004; Kort, Reilly, and Picard, 2001), engaged concentration (see Csikszentmihalyi, 1990), frustration (see Kort et al., 2001; Patrick, Skinner, and Connell, 1993), and other. (In the BROMP coding scheme, behavior categories such as on-task, off-task, or gaming the system are also recorded, but these are not considered in this paper because they are distinct from the affect data.) BROMP coders record the affective state of each student individually, in a pre-determined order that is enforced by the Human Affect Recording Tool (HART) application (Baker et al., 2012; Pardos, Baker, San Pedro, Gowda, and Gowda, 2013) for the Android phone. This strict ordering avoids bias towards interesting or dramatic events in the classroom, ensuring that categories like engaged concentration are accurately represented in the data. Coders have up to 20 seconds to make and verify their assessment, but record only the first affect they identify. In order to build detectors of the affective states identified by the coders, field observations are synchronized with the log files of student interactions with the software. In addition to enforcing BROMP coding scheme, HART synchronizes each observation to within 2 seconds of internet time, allowing us to accurately match each field observation window to the 20-second clip of that student's interactions that are recorded in the software's log file.

Once the two data sources were synchronized, features were distilled based on the information about student interactions with the software that was available in the log files, including temporal features, skill-based features, and features based on the number of errors, the number of correct answers, and the number of hints requested. Many of the 69 features that were ultimately selected for these models were used in recent research, including those previously used in the detection of affect (Baker et al. 2012; Pardos et al., 2013) and in the detection of students who were gaming the system (Baker et al. 2012). Others were derived specifically for this study.

Next, standard data mining classification algorithms (including J48, REPTree, JRip, and K*) were applied within RapidMiner 4.6 (Mierswa, Wurst, Klinkberg, Scholz, and Euler, 2006), a software system that facilitates data mining analyses. These were used to create multiple models for each affective state and population (as well as the models for each affective state created for the combined population).

A rigorous selection process was then applied to determine which model was most appropriate for each affect/population. Detectors for each data set were validated using 5-fold student-level cross-validation. In this process, students are randomly distributed into five equally-sized groups. Detectors are trained on four of the groups while the fifth is held out for testing. Cross-validating at this level assesses the degree to which each model will be effective for new students drawn from the training population. That is, if a detector is trained only on urban students, this type of cross-validation assures that it will perform well for other urban students. It does not assess model goodness for new populations (e.g. how well an urban-trained detector will perform on data from rural students), which will be discussed below.

In conjunction with student-level cross-validation methods, goodness metrics were used to select the optimal models for each category being considered. EDM researchers commonly use multiple metrics during the detector selection process, since no single metric fully captures every aspect of a model. In this study, two goodness (performance) metrics were used to determine which detectors were most effective for each population: Cohen's Kappa (Cohen 1960) and A' (Hanley and MacNeil 1982). Each of these metrics was applied at the level of the clip (the 20-second interval of the log file that forms a basic unit for our analysis).

Cohen's Kappa and A' differ in important ways, but both use very similar scales. Cohen's Kappa assesses the degree to which the detector is better than chance at identifying which clips involve a specific affective state. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. For example, a Kappa of 0.358 would indicate that a detector of boredom is 35.8% better than chance at determining which clips in a large sample included instances of boredom. A' is the probability that the detector will correctly identify whether a specific affective state is present or absent in a specific clip. A' is equivalent to W, the Wilcoxon statistic, and closely approximates the area under the Receiver-Operating Curve (Hanley & MacNeil 1982). A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. For example, an A' of 0.65 would indicate that a detector of boredom can distinguish a bored student from a non-bored student 65% of the time. A' takes detector confidence into account when doing this; for example, if the model thought a bored student had a 62% probability of being bored, while a non-bored student had a 61% probability of being bored, it would give credit to the model for seeing that the bored student had a higher probability of being bored than the non-bored student. In other words, Cohen's Kappa assesses a model's final decisions (and is therefore a better assessment of how well the model will perform when used to drive interventions), while A' assesses a model's confidence in its decisions (and is therefore a better assessment of how well the model will perform when used in discovery with models analyses, which typically take confidence into account).

Once optimal detectors were selected for each population, we tested one against another. For example, the urban-trained detectors of each affective state were tested on suburban and rural populations, and goodness metrics were calculated to determine how effective each detector was at generalizing to the new population. Similarly, the detectors that were trained on the combined data (urban, suburban, and rural) were tested on the individual sub-populations. These tests allow us to compare the degree to which each detector can generalize across the different populations in this study.

Results

Table 1 provides an overview of the best affect detectors for each of the three –sub-populations considered in this study and the combined population. Table 1 also reports the algorithm used to achieve each model, demonstrating the importance of trying multiple algorithms. Details about the 69 features used to construct these models are beyond the scope of the current article, but we can report that there was very little overlap among the different detectors in terms of which features were used for each detector. This was true even when we looked for patterns based on affective states or based on population. For example, the four boredom detectors constructed from different populations (i.e. trained on urban, suburban, rural, or combined data) used different features. Similarly, the features for the four detectors trained on urban data (for boredom, confusion, engaged concentration, and frustration) also showed little to no overlap.

[Insert Table 1 about here]

The results in Table 1 show that all of the detectors trained on sub-populations (urban, suburban, and rural-trained detectors) perform above chance (Kappa > 0.0, A' > 0.5) under student-level cross-validation. Among the individual populations, Kappa ranges from .014

(confusion among rural students) to 0.38 (confusion among suburban students), averaging 0.26; meanwhile A' ranged from 0.56 (confusion among rural students) to 0.74 (confusion among suburban students), averaging 0.67.

Table 1 also presents detectors trained on the combined population. That is, data from urban, suburban, and rural students were combined, then put through the same 5-fold cross-validation process that was applied to individual populations. For the validation of this detector, students were randomly assigned to one of five groups without regard to their urbanicity. On average, the detectors constructed from this combined population performed slightly worse than those trained on single populations (average Kappa = 0.24 and average A' = 0.65), but still well above chance and still within the range found among the detectors trained on individual groups (average Kappa = 0.26 and average A'= 0.665). While these values are far from perfect, they are not substantially below the current state of the art for sensor-free automated detection of student affect (see models and literature reviews in Pardos et al., 2013 and Baker et al., 2012), and detectors of these constructs that perform at this level are capable of predicting standardized exam scores (Pardos et al., 2013) and college attendance (San Pedro, Baker, Bowers, and Heffernan, 2013).

The results for the combined detectors, as shown in Table 1, look promising for researchers who would like to apply their detectors at scale, a task that is easier implement if a single detector can achieve population validity. However, a more complicated story arises if the detectors trained on the combined population are validated according to student urbancity. Tables 2-5 present this information for each affective state, comparing the results of the combined detectors to those trained on individual populations. For comparability, Tables 2-5 repeat the findings in Table 1, giving the student-level, cross-validated goodness metrics for each detector in bold; these numbers are compared to the performance of detectors when applied to populations that are different than the ones they were trained on (e.g. the goodness metrics from the urban-trained detector's performance when applied to the urban population are given in bold, but the same detector's performance is given in regular type-face when applied suburban or rural populations). In other words, the Kappa and A' in Table 2-5 for the cases where the training population and test population are the same (urban→urban, rural→rural, suburban→suburban) reflect model performance when the model is applied to new, unseen students from the same population as the population used in developing the model. In the cases where the training population is a single population and the test population is a different population (urban→rural, urban→suburban, suburban→urban, suburban→rural, rural→urban, rural→suburban), the Kappa and A' reflect model performance when the model is applied to new, unseen students from a different population than the population used in developing the model. And finally, in the cases where the training population is the combined population and the test population is a single population (combined→urban, combined→rural, combined→suburban), the Kappa and A' reflect model performance when the model is applied to new, unseen students from a subset of the population used in developing the model. In all three cases, the model is being tested on new, unseen students, so the comparison between models is fair.

Given the rapidly advancing standards of the field and the steadily improving performance of affective models, it is difficult to select a strict cut-off for importance or significance. Instead, we have marked results that exceed the standards found in other recent publications on affect detection with asterisks. A single asterisk is given for Kappa values exceeding 0.16 (e.g. Sabourin, Mott, and Lester, 2011) and a double asterisk is given for values exceeding 0.30 (e.g. Baker et al., 2012). For A', values of 0.55 are marked with a single asterisk

and values of 0.63 (e.g. Baker et al., 2012) are marked with a double asterisk. These benchmarks represent the changing state of the art in affect detection, based on results achieved in these recent papers (which themselves represented state-of-the-art performance at sensor-free detection with student-level cross-validation, at their respective time of publication).

As is highlighted in Tables 2-5, the detectors of each affective state (boredom, confusion, engaged concentration, and frustration) generally performed above chance (Kappa > 0, A' > 0.5), when tested on new populations. For example, in Table 2 we see that the boredom detector trained on the urban population, which achieves respectable goodness metrics when tested upon new urban students (Kappa = 0.23 and A' = 0.63), achieves a much poorer Kappa of 0.05 when applied to the suburban population (but achieves a comparable A' of 0.62). When applied to the rural population, this urban-trained detector does even worse (Kappa = 0.04 and A' = 0.57).

[Insert Tables 2-5 about here]

In some cases, the detectors performed at or below chance when applied to a new population, although urban and suburban trained detectors appear to be more compatible with each other than with the rural-trained detectors. Take for example the suburban trained detector for engaged concentration in Table 4, which performed at chance according to both goodness metrics when applied to rural students (Kappa = 0.00, A' =0.48) and when applied to urban students (Kappa = 0.02, A' = 0.49). The suburban-trained detector for boredom, shown in Table 2, had more mixed results. When tested on urban students, it achieved a Kappa of .09 (basically chance) but an A' of 0.56 (as good as some other recent research). However, when tested on rural students, both performance metrics indicated that this detector performed below chance (Kappa of -0.03 and an A' of 0.42).

Overall, results suggest that detectors trained on a single sub-population do not generalize well on new population. The average performance under these conditions is Kappa = 0.03 and A' = 0.52, which is perhaps slightly better than chance, but much worse than the original performance of the detectors on the populations they were developed for (average Kappa = 0.26, A' = 0.67). These results indicate that it is generally not safe to apply detectors that have been trained only on one population to data from any of the others, a finding that EDM researchers should pay considerable attention to when choosing research populations.

These results also indicate that the combined detectors, which appeared to be relatively powerful when tested on the combined population using typical EDM cross-validation techniques (average Kappa = 0.24 and average A' = 0.65), may not be equally effective on all populations. Table 6 provides the average performance metrics for detectors trained on urban, suburban, rural and combined populations. It, in conjunction with the data presented in Tables 2-5, show that the models trained on the combined data are, on average, better at detecting affective states in each of the sub-populations than the models trained on one population and applied to another. However, the combined models are more effective at detecting affect among urban and suburban students (Kappa = 0.18, 0.16; A' = 0.62, 0.66) than they are at detecting affect among rural students (Kappa = 0.06; A' = 0.54), where they are just barely better than chance.

[Insert Table 6 about here]

These results suggest that, at least in terms of how students' affect is reflected in their software interaction, urban and suburban students are more similar than suburban and rural

students. These results are quite interesting, since they validate the concerns about ensuring diversity in our research samples, even in terms of categories (like urbanicity) that do not attract as much attention among social scientists as income and ethnic/racial categories. In this case, urbanicity appears to be directly relevant, since the suburban population was comprised mainly of affluent White students and the urban population was comprised mainly of non-White students of lower socioeconomic (SES) status. Both populations were distinct from the rural population in this study, which was comprised mainly of White students from lower SES backgrounds.

Conclusions

One of the unique aspects of EDM and learning analytics analyses, compared to traditional statistics, is that it does not require a pre-conceived hypothesis to be fully formed before processing the data. If it is possible to derive a feature from the log files, it is possible to test the relative effectiveness of this feature in modeling a given construct (e.g. an affective state). This mitigates the risk of confirmation bias that results from only testing a hypothesis stemming from a single theoretical standpoint and expands the opportunities for researchers to explore interactions that might not be easily predictable based on a researcher's own experiences with the data. As such, EDM methods have emerged as a useful paradigm for conducting research on learning within information and communication technologies, where it is now possible to collect large-scale data *in situ*.

As these opportunities arise, so do new challenges for assuring validity. Data collected *in situ* allows us to more capture classroom conditions that are sometimes impossible to model in highly controlled laboratory studies, helping us to ensure the external validity of our models. As more and more of this technology makes it way into the classroom, educators who use these models (either for research or within their own classroom) should be aware that one kind of external validity—population validity—may still be a concern if researchers do not adequately address it during their data collection process.

At the same time, classroom teachers should know that EDM researchers are increasingly rising to these challenges and that they are developing methods to address them. For example, the performance metrics of the models generated in this study, while far from perfect, indicate that the models are adequate for use in of "fail-soft interventions," which are interventions that the software can present to the student that have a low-cost if they are given to a student who does not need them (e.g. presenting a student with encouragement even when we're not entirely certain that they are frustrated.) They are also adequate for generating teacher reports that could help to improve classroom instruction, and an earlier version of these models was used in long-term prediction models, where it was useful in predicting which students are most likely to attend college (San Pedro, Baker, Gowda, and Heffernan. 2013).

In this paper, we have shown that EDM researchers should consider student diversity when constructing automated detectors intended for general use, specifically urbanicity, which is both an important difference between students in its own right and a proxy for ethnicity and socioeconomic status. Specifically, we find that models built using a single population (among urban, rural, and suburban students) do not generalize well to different populations (average Kappa = 0.03; average A' = 0.52). This suggests that care ought to be taken to ensure that other socially significant categories (e.g. SES, gender, ethnicity, and age) are well represented in EDM models, both inter- and intra-culturally.

Importantly, though, we also demonstrate that it is possible to construct a single model that achieves population validity across disparate populations. In this paper, we have improved population validity for detectors of four different affective categories (boredom, confusion, engaged concentration, and frustration) by ensuring that data from all three populations were included during the modeling process and that adequate cross-validation techniques are applied as the optimum algorithms are selected. These detectors were almost as effective on the combined population (average Kappa = 0.24; average A' = 0.65) as the more customized detectors were on the populations they were modeled after (average Kappa = 0.26; average A' = 0.67), and have the added advantage of being applicable at a much greater scale. However, the combined detectors were most successful at finding commonality between the urban and suburban populations. The combined detectors were less successful when applied to the rural population, where they performed only slightly better than the models generated using only the suburban or the urban population.

Our findings have an important consequence: they suggest that automated detectors of affect cannot be safely applied to a new population unless unless researchers have used similar populations to develop them. Since EDM and learning analytics research often aim to improve both automated and teacher-led interventions, these results demonstrate the importance of verifying population validity before applying these interventions at scale. It does no good to prompt a teacher that a student is experiencing extended boredom, for example, if that prompting is based on invalid model.

These findings complement a small but growing body of research on population validity in EDM research that also has important consequences for those who are more generally interested in cross-cultural effects on the educational process. However, previous research has shown that, unlike these affect models, models of some constructs do generalize to quite different populations. For example, models of carelessness and effective help-seeking created using data from either America and the Philippines have been found to function effectively in the other country (San Pedro et al., 2011) even though the social and educational contexts are quite different between the two countries. By contrast, however, Soriano et al., (2012) shows that help-seeking models that generalize between the USA and Philippines cannot be safely applied to students in Latin America. It is possible that affective states are particularly susceptible to cultural variation, heightening the need for tests of population validity when they are the constructs being investigated. As such, affect appears to differ between cultures more than help-seeking or carelessness, a finding worthy of future investigation. More extensive qualitative analyses of both the features identified in the log files and the experiences of field observers could help us to identify what it is about these constructs that makes them sensitive to cultural influences.

Further research is needed to establish which constructs are vulnerable to cultural differences, but also which inter- and intra-cultural differences matter. In this study, we have shown that urban, suburban, and rural students from the same region of the United States are sufficiently different from one another that we cannot expect an affect detector based on one population to adequately generalize to another. Future research may show that this particular intra-national difference is unique to rather wealthy and industrialized countries (i.e. those that Blanchard complains that the field relies too heavily upon), since urbanicity in this country might have a greater effect on access to educational technology than it does in countries where even the wealthier areas have rather limited technological access. If this is the case, then the inclusion of

more vulnerable US populations in EDM research may help to improve the validity of our models when applied overseas.

By increasing model validity to broader populations, we ensure that discovery with models research and interventions based on these types of automated detectors have the maximum potential applicability. In recent years, just within the context of ASSISTments, several discoveries have been made about affect using automated detectors, including how affect is shaped by knowledge (San Pedro et al., 2013a), how interface design influences affect (Hawkins et al., 2013), how affect influences learning (Pardos et al., 2013), and how affect influences the eventual decision to attend college (San Pedro et al., 2013b). Verifying that these findings apply to a broad population of learners is essential to using these findings to drive educational practice.

In general, there is much to understand about the populations that use information and communication technologies for education, and we will need to understand these factors for ICT-based education to reach its full potential. As technological access expands, EDM researchers will have more and more opportunities to explore such cultural differences and to contribute to educators' understanding of the learning process.

Acknowledgments

We would like to thank the Bill & Melinda Gates Foundation (#OPP1048577) and the National Science Foundation (#DRL-1031398) for their support for this research. Thanks also to the editors and anonymous reviewers who offered very thoughtful advice on our revisions. Any errors are, of course, our own.

References

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., & Rossi, L. (2012). Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126-133.

Blanchard, E. G. (2012). Intelligent Tutoring Systems On the WEIRD Nature of ITS/AIED Conferences. *Proceedings of the 11th International Conference on Intelligent Tutoring Systems*, 280-284.

Campbell, N. J. (1989). Computer anxiety of rural middle and secondary school students. *Journal of Educational Computing Research, 5*, 2, 213-220.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1, 37- 46.

Chen, R. S. & Liu, I. F. (2013). Research on the effectiveness of information technology in reducing the Rural–Urban Knowledge Divide. *Computers & Education*, *63*, 437-445.

Craig, S.D., Graesser, A.C., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media,* 29, 3, 241-250.

Csikszentmihalyi, M., (1990). *Flow: The Psychology of Optimal Experience.* Harper-Row, New York.

D'Mello, S.K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T. & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18, 1-2, 45-80.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4,* 5-6, 304-317.

Graham, S. E. & Provost, L. E. (2012). Mathematics achievement gaps between suburban students and their rural and urban peers increase over time. Issue Brief No. 52. *Carsey Institute*. URL http://files.eric.ed.gov/fulltext/ED535962.pdf

Hanley, J. & McNeil, B. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143, 29-36.

Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Wixon, M. & Sao Pedro, M. (2013). Discovery with Models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist, 57,* 10, 1479-1498.

Hu, S. (2003). Educational aspirations and postsecondary access and choice: Students in urban, suburban, and rural schools compared. *Education Policy Analysis Archives*, 11, 14, 14.

Johnson, D. L. (2005). Computer tutors get personal. *Learning and Leading with Technology*, 33, 3, 14.

Kort, B., Reilly, R. & Picard, R. (2001). An Affective Model Of Interplay Between Emotions And Learning: Reengineering Educational Pedagogy—Building A Learning Companion. *Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges, Madison, Wisconsin: IEEE Computer Society*, 43-48.

Markauskaite, L., Kennan, M.A., Richardson, J., Aditomo, A. & Hellmers, L. (2012). Investigating eResearch: Collaboration Practices and Future Challenges. In A.A. Juan, T. Daradoumis, M. Roca, S.E. Grasman, & J. Fauli (Eds.) *Collaborative and Distributed E-Research: Innovations in Technologies, Strategies, and Applications.* Hershey, PA: IGI Global.

Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. *Journal of Research on Computing in Education*, *41*, 3, 331.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, 935-940.

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, *88*, 2, 203.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E. & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*, 4, 1, 11-48.

Ogan, A., Walker, E., Baker, R.S.J.d., de Carvalho, A., Laurentino, T., Rebolledo-Mendez, G. & Castro, M.J. (2012). Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. *Proceedings of ACM SIGCHI: Computer-Human Interaction*, 1381-1390.

Paez, D. & Vergara, A.I. (2000). Theoretical and Methodological Aspects of Cross-Cultural Research. *Psicothema*, 12, 1-5.

Pardos, Z.A., Baker, R.S.J.d., 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*, 117-124.

Patrick, B. C., Skinner, E. A. & Connell, J. P. (1993). What motivates children's behavior and emotion? Joint effects of perceived control and autonomy in the academic domain. *Journal of personality and social psychology*, *65*, 4, 781.

Rappa, N. A., Yip, D. K. H., & Baey, S. C. (2009). The role of teacher, student and ICT in enhancing student engagement in multiuser virtual environments. *British Journal of Educational Technology*, *40*, 1, 61-69.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A. & Rasmussen, K.P. (2005). The Assistment project: Blending assessment and assisting. *Proceedings of the 12th Artificial Intelligence in Education*, 555-562.

Reimann, P., Thompson, K., & Weinel, M. (2007). Collaborative learning by modelling: Observations in an online setting. *ICT: providing choices for learners and learning. Proceedings of Ascilite Singapore*, 887-897.

Rupp, A.A., Gushta, M., Mislevy, R.J. & Shaffer, D.W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment, 8*, 4, 4-47.

Rupp, A. A., Levy, R., DiCerbo, K., Sweet, S., Crawford, A. V., Caliço, T., Benson, M., Fay, D., Kunze, K.L., Mislevy, R.J., & Behrens, J.T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, *4,* 1, 49-110.

Sabourin, J., Mott, B. & Lester, J. C. (2011). Modeling learner affect with theoretically grounded dynamic bayesian networks. In *Affective Computing and Intelligent Interaction* (pp. 286-295). Berlin & Heidelberg: Springer.

San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J. & Heffernan, N.T. (2013a). Predicting college enrollment from student interaction with an Intelligent Tutoring System in middle school. *Proceedings of the 6th International Conference on Educational Data Mining*, 177-184.

San Pedro, M.O.Z., Baker, R.S.J.d., Gowda, S.M. & Heffernan, N.T. (2013b). Towards an understanding of affect and knowledge from student interaction with an Intelligent Tutoring System. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 41-50.

San Pedro, M.O.C., Baker, R. & Rodrigo, M.M. (2011). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 304-311.

Sao Pedro, M., Baker, R.S.J.d. & Gobert, J. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)*, 249-260.

Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior, 25*, 3, 167-185.

Siemens, G. & Baker, R.S.J.d. (2012). Learning Analytics and Educational Data Mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. 252-254.

Smeets, E. & Mooij, T. (2001). Pupil- centred learning, ICT, and teacher behaviour: observations in educational practice. *British Journal of Educational Technology*, *32*, 4, 403-417.

Wenglinsky, H. (1998). Does it Compute? The Relationship Between Educational Technology and Student Achievement in Mathematics.

Wilson, K. R., Wallin, J. S. & Reiser, C. (2003). Social stratification and the digital divide. *Social Science Computer Review*, *21*, 2, 133-143.

Tables and Figures

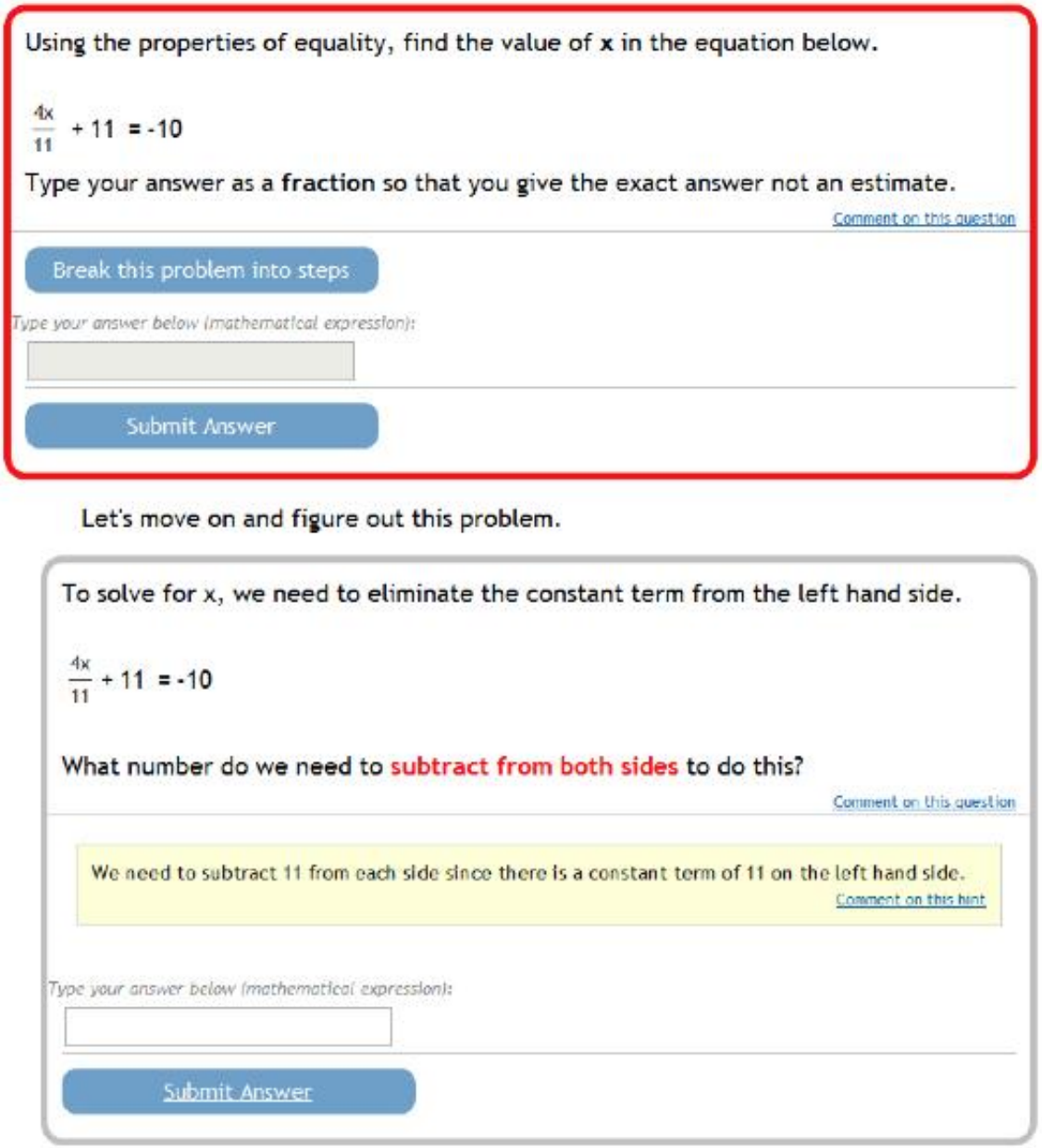

Figure 1: Scaffolded Mathematics Instruction in ASSISTments

Table 1

Summary of model description and cross-validated performance

Detector	Algorithm	Kappa	A'
Urban-trained boredom detector	JRip	0.23	0.63
Urban-trained confusion detector	J48	0.27	0.74
Urban-trained engaged concentration detector	K*	0.36	0.68
Urban-trained frustration detector	REPTree	0.29	0.74
Suburban-trained boredom detector	REPTree	0.19	0.67
Suburban-trained confusion detector	REPTree	0.38	0.74
Suburban-trained engaged concentration detector	J48	0.27	0.63
Suburban-trained frustration detector	REPTree	0.17	0.59
Rural-trained boredom detector	K*	0.24	0.73
Rural-trained confusion detector	JRip	0.14	0.56
Rural-trained engaged concentration detector	REPTree	0.37	0.71
Rural-trained frustration detector	JRip	0.20	0.57
Boredom detector trained on combined population	J48	0.24	0.66
Confusion detector trained on combined population	REPTree	0.15	0.63
Engaged Concentration detector trained on combined population	JRip	0.42	0.73
Frustration detector trained on combined population	JRip	0.15	0.60

Table 2

Comparison of boredom detectors performance across sub-populations

	Urban		Suburban		Rural	
	Kappa	A'	Kappa	A'	Kappa	A'
Urban trained	**0.23***	**0.63****	0.05	0.62*	0.04	0.57*
Suburban trained	0.09	0.56*	**0.19***	**0.67****	-0.03	0.42
Rural trained	0.02	0.53	0.02	0.56*	**0.24***	**0.73****
Trained on combined population	0.10	0.55*	0.20*	0.72**	-0.02	0.53*

Note: Values of $-1 \leq \text{Kappa} < 0$ and $0 \leq A' < 0.5$ indicate model performance below chance, while values of $0 < \text{Kappa} \leq 1$ and $0.5 < A' \leq 1$indicate performance that is better than chance. Numbers marked with asterisks indicate performance that meets or exceeds benchmarks set by recent EDM publications.

Table 3

Comparison of confusion detectors performance across sub-populations

	Urban		Suburban		Rural	
	Kappa	A'	Kappa	A'	Kappa	A'
Urban trained	**0.27***	**0.74****	0.03	0.37	0.00	0.51
Suburban trained	-0.01	0.51	**0.38****	**0.74****	-0.02	0.51
Rural trained	-0.03	0.51	-0.02	0.52*	**0.14**	**0.56***
Trained on combined population	0.22*	0.67**	0.10	0.66**	0.02	0.52

Note: Values of -1 ≤ Kappa < 0 and 0 ≤ A'< 0.5 indicate model performance below chance, while values of 0 < Kappa ≤ 1 and 0.5 < A'≤ 1indicate performance that is better than chance. Numbers marked with asterisks indicate performance that meets or exceeds benchmarks set by recent EDM publications.

Table 4

Comparison of engaged concentration detectors performance across sub-populations

	Urban		Suburban		Rural	
	Kappa	A'	Kappa	A'	Kappa	A'
Urban trained	**0.36****	**0.68****	0.07	0.58*	0.18*	0.58*
Suburban trained	0.06	0.47	**0.27***	**0.63****	0.00	0.48
Rural trained	0.16*	0.42	0.12	0.61*	**0.37****	**0.71****
Trained on combined population	0.21*	0.64**	0.34**	0.66**	0.27*	0.62*

Note: Values of -1 ≤ Kappa < 0 and 0 ≤ A'< 0.5 indicate model performance below chance, while values of 0 < Kappa ≤ 1 and 0.5 < A'≤ 1indicate performance that is better than chance. Numbers marked with asterisks indicate performance that meets or exceeds benchmarks set by recent EDM publications.

Table 5

Comparison of frustration detectors performance across sub-populations

	Urban		Suburban		Rural	
	Kappa	A'	Kappa	A'	Kappa	A'
Urban trained	**0.29***	**0.74****	0.00	0.45	-0.02	0.52
Suburban trained	0.02	0.49	**0.17***	**0.59***	-0.04	0.52
Rural trained	0.01	0.55*	-0.01	0.55*	**0.20***	**0.57***
Trained on combined population	0.18*	0.61*	-0.01	0.60*	-0.02	0.49

Note: Values of -1 ≤ Kappa < 0 and 0 ≤ A'< 0.5 indicate model performance below chance, while values of 0 < Kappa ≤ 1 and 0.5 < A'≤ 1indicate performance that is better than chance. Numbers marked with asterisks indicate performance that meets or exceeds benchmarks set by recent EDM publications.

Table 6

Average detector performance (all affective states) by population

	Urban		Suburban		Rural	
	Kappa	A'	Kappa	A'	Kappa	A'
Urban trained	**0.29***	**0.70****	0.04	0.50	0.05	0.54
Suburban trained	0.04	0.51	**0.25***	**0.66****	-0.02	0.48
Rural trained	0.04	0.50	0.03	0.56*	**0.24***	**0.64****
Trained on combined population	0.18*	0.62*	0.16*	0.66**	0.06	0.54

Note: Values of $-1 \leq \text{Kappa} < 0$ and $0 \leq A' < 0.5$ indicate model performance below chance, while values of $0 < \text{Kappa} \leq 1$ and $0.5 < A' \leq 1$indicate performance that is better than chance. Numbers marked with asterisks indicate performance that meets or exceeds benchmarks set by recent EDM publications.