Role of Socio-Cultural Differences in Labeling Students' Affective States

Eda Okur[1], Sinem Aslan[1], Nese Alyuz[1], Asli Arslan Esme[1], and Ryan S. Baker[2]

[1] Intel Labs, Intel Corporation, Hillsboro OR, USA
{eda.okur, sinem.aslan, nese.alyuz.civitci,
asli.arslan.esme}@intel.com
[2] University of Pennsylvania, Philadelphia PA, USA
rybaker@upenn.edu

Abstract. The development of real-time affect detection models often depends upon obtaining annotated data for supervised learning by employing human experts to label the student data. One open question in labeling affective data for affect detection is whether the labelers (i.e., human experts) need to be socio-culturally similar to the students being labeled, as this impacts the cost and feasibility of obtaining the labels. In this study, we investigate the following research questions: For affective state labeling, how does the socio-cultural background of human expert labelers, compared to the subjects (i.e., students), impact the degree of consensus and distribution of affective states obtained? Secondly, how do differences in labeler background impact the performance of affect detection models that are trained using these labels? To address these questions, we employed experts from Turkey and the United States to label the same data collected through authentic classroom pilots with students in Turkey. We analyzed within-country and cross-country inter-rater agreements, finding that experts from Turkey obtained moderately better inter-rater agreement than the experts from the U.S., and the two groups did not agree with each other. In addition, we observed differences between the distributions of affective states provided by experts in the U.S. versus Turkey, and between the performances of the resulting affect detectors. These results suggest that there are indeed implications to using human experts who do not belong to the same population as the research subjects.

Keywords: Affective State Labeling, Student Affect Detection, Cross-Cultural, Inter-Rater Agreement, Intelligent Tutoring Systems (ITS).

1 Introduction

Automated detection of learner affect has matured as an area of Artificial Intelligence in Education (AIED) research, with models of learner affect now published for a range of learning environments [1-3]. These models have formed the basis of a range of scientific analyses, including the relationship between affect and student outcomes [3], and the differences in learning outcomes between brief and extended affect [4]. They have also been used as the basis for automated interventions which improve students'

affect and engagement, and in turn their learning, by responding to negative affect in real time [5, 6].

These automated detectors of affect are typically created through a supervised learning approach, where the first step is to collect some kind of external, ground-truth measure of a student's affect at specific points in time, whether those labels are from self-report, video coding, or field observation. These ground-truth labels are often useful scientific interests in their own right, and are used for analyses such as understanding the dynamics of student affect over time [7], and the affect that students experience within MOOCs [8].

Early work on the expert coding of affect and emotion, such as the Facial Action Coding System (FACS) focused on deriving rationally understandable and widely-agreed, culturally universal indicators of affect [9]. However, the inter-rater reliability for this approach was poor [10]. More recent approaches to the expert coding of affect have relied on more subjective judgments. Two such approaches, the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [11] and the Human Expert Labeling Process (HELP) [12], have each achieved inter-rater agreement measures over 0.6, indicating considerably higher levels of agreement than FACS. Neither BROMP nor HELP makes claims to cultural universality; in fact BROMP has been re-normed in four different countries and its coding manual explicitly warns against using raters from a different national origin than the students whose affect is being coded [11]. However, the empirical basis for this caution remains insufficient, and many affective computing research groups continue to use raters from a different national origin than the subjects they are studying, both for AIED research and within other application areas. Even with BROMP, there are anecdotal reports of individuals successfully achieving acceptably high inter-rater reliability outside their original home country, after living in a different country for several years [11].

If it was possible to use the same affect labeling protocol internationally without re-norming, wide use of affect labeling would become considerably more feasible. In addition, if it were feasible to use raters of any national origin, it would become more feasible to conduct affective computing research and development even in countries where protocols have already been normed.

As such, this paper investigates the degree to which human expert labelers from the same country as the students being labeled actually achieve higher inter-rater reliability than experts from a different country, when labeling the affective states of learners using an online platform for mathematics. In doing so, we analyze both the within-country and cross-country agreements in affect labeling, as it is possible that experts from a different country than students may agree with each other but disagree with experts from the students' country, a pattern that would suggest systematic error and bias in affect labeling.

2 Methodology Overview

2.1 Research Questions

In this research, we investigate whether human experts who are from the same country as the subjects (i.e., students) would provide different affective state labels than the ones who are from another country. We hypothesize that such difference could be attributed to the fact that same-culture experts will be familiar with the learners' context (going through the same schooling experience, being familiar with the learning context) and culture (sharing the similar values and conceptions about affect in learning, as well as similar modes of expressing affect). Hence, they could provide more reliable affect labels. Towards this end, this study investigates the following research questions: How is the socio-cultural background of human expert labelers, compared to the students, associated with the degree of consensus and distribution of affective states labeled, for the task of affective state labeling? Secondly, how would the differences in labelers' background impact the performance of affect detection models that are trained using these labels?

2.2 Data Collection and Labeling

Research Context and Data Collection. We collected data from 9th-grade students (ages 14-15) through authentic classroom pilots in an urban high-school in Turkey. The pilots took place in an optional Math course offered during school hours throughout a school semester. 13 pilot sessions of 40 minutes each were provided to 17 volunteering students; around 113 hours of student data is collected in total. During the pilots, the students used an online learning platform where they watched instructional videos (i.e., *Instructional* sections) and solved objective assessments to test their mastery (i.e., *Assessment* sections). A Math teacher participated in the course as a facilitator of the learning process – i.e., whenever the students needed her input, she got involved. The curriculum of the course was designed in collaboration with the course teacher. The data for each student was collected individually, using a laptop equipped with a camera. While the students were involved in learning activities on the online platform, we collected two streams of videos: (1) videos of the student from the camera, to enable monitoring of observable cues available in the individual's face or upper body; and (2) student desktop videos, to observe contextual information from the learning activity. We also recorded system interaction logs and whether students were participating in instructional or assessment activities.

Data Labeling. Using the Human Expert Labeling Process (HELP) [12], detailed below, we had five human experts from Turkey and five human experts from the United States with at least a B.S. degree in Psychology/Educational Psychology label the same portion of the student data collected from the pilots. Each group of human experts labeled around 14 hours of data collected from ten students in two sessions. The experts defined segments based on observed state changes (i.e., an expert defined segments based on identifying a change in affect rather than using pre-defined segments of pre-

defined length) and provided labels with regards to affective states during learning. These experts labeled student data after receiving face-to-face, instructor-led training provided by the research team. The training involved a presentation explaining operational definitions of labels and providing examples, as well as a practice session with the labeling tool. Note that the training was not prescriptive in that we did not give directive instructors such as telling the experts that the student's fist on their cheek indicates boredom. This is because annotating affective states is a highly subjective task and we expected the human experts to infer emotional states just as teachers would do in classrooms. This more subjective coding approach is similar to what is seen in BROMP [11], and distinct from FACS [13], which is more prescriptive. The research team provided the same training materials for operational definitions of these labels within both countries (translated), and attempted to use similar training procedures, though there remained some slight differences in the actual training process due to differences between the experts and the settings where they were being trained. The definitions of emotional states provided to the human experts are given in Table 1.

Table 1. Operational definitions of emotional states.

Satisfied	If a student is not having any emotional challenges during a learning task. This can include all positive emotional states of the student from being neutral to being excited during the learning task.
Bored	If the student is feeling bored during the learning task.
Confused	If the student is getting confused during the learning task – in some cases this state might include some other negative states such as frustration (which can be viewed as an increased level of confusion).

In addition to these emotional states, the experts also had two other labels: *Can't Decide* (i.e., if the expert cannot decide on the final emotional state) and *N/A* (i.e., if the data cannot be labeled - e.g., there is no one in front of the camera). Using the HELP Labeling Tool (see Fig. 1 for a sample view), the experts annotated the data using all available cues, such as video and audio capture of the student, desktop recording with mouse cursor locations emphasized, and contextual logs from the device and content platform. Both of the human expert groups completed labeling on the same student data using the same labeling tool; the labeling itself took place at different times and locations.

Human Experts' Demographics. The demographics for the human experts in Turkey were rather different than the ones in the United States. The age-range was 20s-30s in the former group, whereas in the latter it ranged from 30s-60s. All the experts in Turkey were female whereas we had one male expert in the United States. All experts had a Psychology/Educational Psychology degree in both groups and some of them had experience as a classroom teacher.

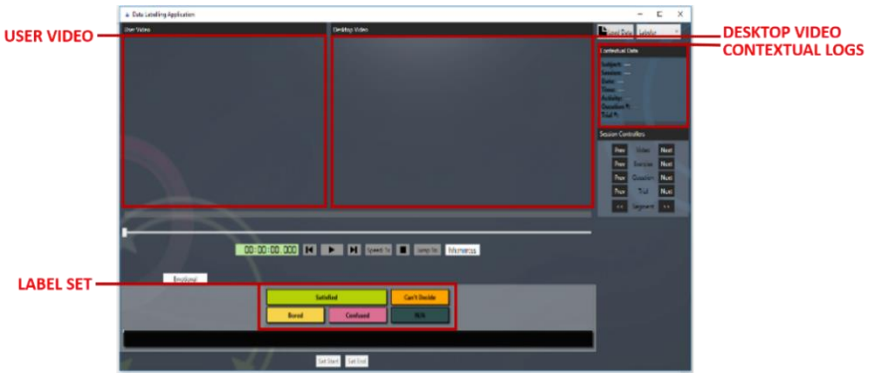

Fig. 1. HELP Labeling Tool (sample view).

2.3 Data Analysis

Data Pre-processing. In order to obtain the inter-rater agreement measures, the labeled data was pre-processed as follows: First of all, we applied data aggregation (i.e., creating instances by sliding-windows) on the labeling outputs of all the experts. The labeling data provided by each expert were pre-processed and divided into instances, where for each 8-seconds sliding-window (with an overlap of 4-seconds), an emotional label was assigned by each expert. The section type information is taken into account while creating these windowed data, so that none of the instances contained both instructional and assessment activities. These instance labels were employed to calculate the inter-rater agreement measures.

To compute each affective state's proportion labeled, we distilled a single label (per population of experts) for each time window, which served as a final ground truth label for each group of human experts. In order to do this, we applied majority voting across all five experts within each group. For each data instance, a majority label (i.e., label with at least three-out-of-five votes) was obtained. We also applied majority voting across the best three experts (i.e., three-out-of-five labelers with the highest consensus) within each group, to eliminate the possible influence of an outlier on final ground truth labels for each group. These final majority labels were utilized to train our supervised models for affect detection.

Metrics for Analysis. We calculated the inter-rater agreement among multiple experts to investigate the effect of the socio-cultural background of human experts on the degree of agreement achieved for the affective state labeling. For inter-rater agreement, we used consensus measures which are designed to estimate the degree of agreement among multiple raters [14, 15]. In this study, we used Krippendorff's alpha [16], because it is both suitable for multiple raters and robust for missing rates: It provides a corrected consensus estimate by comparing the observed agreement to the expected agreement (i.e., the agreement that would be obtained by chance alone) [16]. As summarized in [17], there is no gold standard on the interpretation of Krippendorff's alpha

values, and different thresholds are utilized throughout the literature. However, [17] states that values above 0.4 are often considered to represent moderate agreement.

Methods for Analysis. To investigate differences between the human experts from Turkey and the U.S., inter-rater agreement measures were calculated separately in several ways.

First, we calculated within-country agreements. When doing so, we considered both all five experts, and the best three experts of each group, to mitigate against poor results due to outlier experts. Note that "best" refers to the three experts having the highest consensus among all possible three-out-of-five combinations of experts within each group. For the pairwise agreement cases, agreement is computed for each pair of experts, and then averaged across all possible pairs within the corresponding set of experts. We calculated:

- Inter-rater agreement of all five U.S. experts ("US-all-5").
- Pairwise agreement of all five U.S. experts ("US-all-5-pairwise").
- Inter-rater agreement of all five Turkey experts ("TR-all-5").
- Pairwise agreement of all five Turkey experts ("TR-all-5-pairwise").
- Inter-rater agreement of the best three U.S. experts, where "best" is calculated as the highest agreement between three raters (among all combinations of five raters) in the same country ("US-best-3").
- Pairwise agreement of the best three U.S. experts ("US-best-3-pairwise").
- Inter-rater agreement of the best three Turkey experts, where "best" is calculated as the highest agreement between three raters (among all combinations of five raters) in the same country ("TR-best-3").
- Pairwise agreement of the best three Turkey experts ("TR-best-3-pairwise").

Next, we analyzed agreement involving two groups of experts, mixed between both countries, to see how well they agreed with each other, on the whole. Calculating mixed-country group agreements gives us an idea of what the result might be if a research group hired a set of human experts with various backgrounds. We calculated:

- Inter-rater agreement of all ten experts ("Both-all-10").
- Pairwise agreement of all ten experts ("Both-all-10-pairwise").
- Inter-rater agreement of the best three experts from the U.S. and the best three experts from Turkey ("Both-best-6").
- Pairwise agreement of the best three experts from the U.S. and the best three experts from Turkey ("Both-best-6-pairwise").

Finally, we compared between experts in different countries to see how well they agreed with each other, solely in cross-country comparisons. Here, each U.S. expert is compared to each Turkey expert, and the results are averaged across all such pairs. We calculated:

- Pairwise agreement of all ten experts, between cross-country pairs ("Intercountry-all-10-pairwise").

- Pairwise agreement of the best three experts from the U.S. and the best three experts from Turkey, between cross-country pairs ("Intercountry-best-6-pairwise").

By conducting these comparisons, we aimed to understand how much agreement can be achieved within a culture, and how agreement would change when comparing experts from different cultures. Note that *N/A* labels are simply treated as missing values in all of our inter-rater agreement calculations, to avoid misleading higher agreements due to this relatively objective label for which most experts agree on. As a result, we computed inter-rater agreements for the other four labels (i.e., *Satisfied*, *Bored*, *Confused*, *Can't Decide*).

In addition to examining the degree of agreement between experts, we also computed summary statistics on the overall prevalence of affective states labeled by each group, in order to see whether one group of experts generally identified specific affective states more than others (potentially indicating some degree of bias in observation among the U.S. experts, who were labeling students from a different country). We calculated these proportions for all the content taken together (i.e., *Overall*), as well as for each of the system's two types of content (i.e., *Instructional* and *Assessment*), taken individually. These proportions were calculated on the final ground truths computed by majority voting of best three experts of each group separately, again to avoid the influence of an outlier within each group. Note that after this point, we are solely interested in the affective states (i.e., *Satisfied, Bored, Confused*) so we no longer consider *Can't Decide* final labels for either proportions or models of affective states.

As a last step, we explored whether there were differences in the performance of affect detection models trained using final majority labels obtained by each group of expert labelers. In this study, we considered two modalities for affect detection as follows: (1) *Appearance*: upper-body information of students from the camera, (2) *Context & Performance* (C&P): interaction and performance logs of students from the online learning platform. For feature extraction, in order to obtain instance features for each modality, we utilized the same 8-seconds sliding window (with an overlap of 4-seconds) approach that we used to derive the instance labels. For *Appearance*, the raw video data are segmented into instances and time series analysis methods were utilized to extract appearance features, consisting of motion and energy measures (e.g., trend of pose energy), robust statistical estimators (e.g., trimean) of head velocity, and frequency domain features related to head position, pose, and facial expressions. More details of the *Appearance* modality can be found in our previous study [18] where we used the same appearance features in this study. For C&P, we extracted features for inferring affect from the raw user interaction logs collected from the content platform together with the user profile (e.g., gender), consisting of features related to time (e.g., video duration, time from beginning, time spent on attempts/questions, etc.), student performance (e.g., success/failure of attempts, percentage of attempts correct, etc.), attempts made (trial number, number of attempts taken until success, etc.), hints (number of hints used on attempts/questions, etc.), and others (question number, etc.). Some of these extracted C&P features are adapted from the study [3] selecting the subset which are applicable to the content platform we used. More details of the C&P features employed in this study can be found in our previous study [19]. In total, we extracted 188

appearance and 24 C&P features per each instance, which are fed into separate classifiers as feature vectors along with the ground truth labels. For both modalities, we trained Random Forest classifiers with 100 trees. For model training and testing, the labeled datasets (14 hours of data collected from ten students) are partitioned into training (80%) and test (20%) sets, stratifying to keep the distribution of each state and student similar within each group. We applied leave-one-subject-out cross-validation to prevent overfitting, where for each test student, the training samples of all other students were used to construct subject-specific training sets. Due to the class imbalance problem, we employed 10-fold random sample selection to obtain balanced training sets. We compared the performance results of affect detection models trained on the same feature sets but different label sets (i.e., ground truths). In each case, we trained and tested on the final majority labels obtained from the best three experts from the U.S. versus the best three experts from Turkey.

3 Experimental Results

3.1 Results: Inter-Rater Agreement

Inter-rater agreement values among the human experts for the several comparison sets given in Section 2.3.2 are outlined in Table 2. The results of within-country agreements with all five experts given in Table 2 show that when comparing between the experts from the U.S. and Turkey as two separate groups, the experts from Turkey perform moderately better in terms of inter-rater reliability. The results for within-country agreements with the best three experts (discarding the worst, potentially outlier, experts), given in Table 2, is similar although the difference between the two groups decreases. Note that when eliminating the outliers, the improvement we achieve is higher for the U.S. experts than the experts from Turkey, which suggests that the outliers had a more negative impact on the agreement for the U.S. experts.

At first, the within-country results with the best three experts given in Table 2 might look promising, suggesting that the better raters among the U.S. experts perform almost as well as the Turkey experts. However, the mixed-country results given in Table 2 show that when we combine these two groups and obtain a mixed-cultural set, the agreement scores go down, even for the best experts within each group.

Finally, we can look at how closely the two groups agree directly with each other, by comparing pairs of experts (one from the U.S. and one from Turkey), shown as the cross-country results in Table 2. These results indicate that pair-wise cross-cultural comparison has the lowest degree of agreement of any of the comparisons conducted here, reaching a value of 0.4 or lower. Such results might signify that although the experts from the same country could agree with each reasonably well, their agreement drops when comparing their labels with another group from a different country. In a way, this could be the worst possible result for using experts from a different country than participants – it suggests that experts may be systematically biased in their evaluations of student affect, agreeing on a label that may not actually reflect the student's affect. We investigate this possibility in greater depth in the following sections.

Table 2. Inter-rater agreement (Krippendorff's Alpha) measures among human experts from the United States (US) and Turkey (TR).

Human Expert Comparison Sets	Krippendorff's Alpha	Human Expert Comparison Sets	Krippendorff's Alpha
Within-country (all-5)		Mixed-country	
US-all-5	0.469	Both-all-10	0.452
US-all-5-pairwise	0.472	Both-all-10-pairwise	0.446
TR-all-5	0.578	Both-best-6	0.483
TR-all-5-pairwise	0.585	Both-best-6-pairwise	0.478
Within-country (best-3)		Cross-country	
US-best-3	0.560	Intercountry-all-10-pairwise	0.379
US-best-3-pairwise	0.564	Intercountry-best-6-pairwise	0.400
TR-best-3	0.617		
TR-best-3-pairwise	0.626		

3.2 Results: Overall Proportions of Affect

There are also differences in the proportions of labels provided by human experts in the U.S. versus Turkey, as shown in Fig. 2. As Fig. 2 demonstrates, although both of the groups labeled *Confused* at a similar frequency (18.0% vs. 18.7%), the experts in the U.S. thought that students were *Satisfied* almost twice more frequently as the experts in Turkey (55.5% vs. 29.2%). For *Instructional* content, the difference in the *Satisfied* state distribution is even greater (49.1% vs. 18.1%). Similarly, for the *Assessment* content, although the Turkey experts thought that students were *Bored* fairly frequently (27.7%), the U.S. experts considered those students to be *Bored* substantially less frequently (7.7%). The U.S. experts instead annotated the majority of these students as *Satisfied* (60.5%) while solving questions. Note that neither group of experts identified students as *Confused* during the *Instructional* activities.

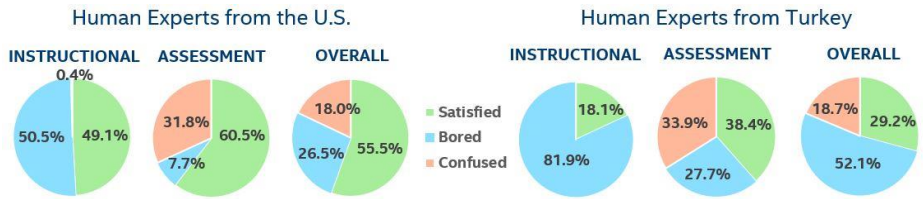

Fig. 2. Students' emotional-state distributions for different section types (i.e., *Instructional*, *Assessment*, and *Overall*) as labeled by the human experts from the U.S. and Turkey.

3.3 Results: Performance of Affect Detection Models

Finally, we compared whether using ground truth labels from same-culture expert labelers (from Turkey) produced better affect detectors than ground truth labels from

cross-culture expert labelers (from U.S.). For each modality (i.e., Appr: *Appearance* and C&P: *Context & Performance*), we trained separate models for different activities (i.e., *Instructional* and *Assessment*), and trained separate models for each group of experts: U.S. and Turkey. The generic (i.e., subject-independent) results for each of these two modalities (Appr and C&P) and activity types (*Instructional* and *Assessment*) are reported in Table 3, broken out by each affective state and by labeler background. We also report the average training set sizes (i.e., average of ten students, where balanced training sets differ for each student due to the leave-one-subject-out methodology) and total test set sizes (i.e., total of ten students' unbalanced test sets).

Note that no students were labeled as *Confused* during *Instructional* activities, so our analysis of the *Instructional* models consists solely of examples of *Satisfied* and *Bored*. For *Assessment* models, although we have relatively balanced classes among the Turkey experts, the U.S. experts seldom labeled students as *Bored* (7.7%), too small a sample to develop detectors on. As our goal here is to compare models between same-culture and cross-culture labels, we trained our *Assessment* models to detect *Satisfied* and *Confused* states only, for both groups of experts.

In Table 3, we report mean F1-scores which are computed by weighted averaging over all folds (10-fold for random selection; times ten for all students), where weights are the test counts of each model. These mean F1-scores are computed for each class (i.e., affective state) and for overall classification performance. In addition to F1, mean AUC (i.e., the area under the ROC curve) values are reported in Table 4, which are again computed by weighted averaging over all folds. We report single AUC value for each model, which reflects the overall performance of binary classifiers.

Table 3. Affect detection classifier results (F1-scores) for separate modalities (Appr: *Appearance*, C&P: *Context & Performance*) and different section types (Instr: *Instructional*, Assess: *Assessment*) trained using labels by experts from the United States and Turkey.

Section Type	Class	Labels: Experts from the U.S.				Labels: Experts from Turkey			
		Avg. Train Count	*Total Test Count*	Appr F1	C&P F1	*Avg. Train Count*	*Total Test Count*	Appr F1	C&P F1
Instr.	Satisfied	*888*	*265*	0.62	0.58	*336*	*94*	0.41	0.42
	Bored	*888*	*271*	0.67	0.59	*336*	*425*	0.86	0.88
	OVERALL	*1776*	*536*	0.65	0.58	*672*	*519*	0.77	0.80
Assess.	Satisfied	*787*	*416*	0.59	0.80	*769*	*243*	0.43	0.73
	Confused	*787*	*219*	0.45	0.63	*769*	*215*	0.57	0.66
	OVERALL	*1574*	*635*	0.53	0.74	*1538*	*458*	0.51	0.70

As shown in Table 3, there are notable differences in the affect detection performances when models are trained using labels provided by experts in the U.S. versus Turkey. For *Instructional* models, both modalities had higher F1-scores for the overall performance when labels provided by Turkey experts are utilized instead of the U.S. experts (US vs. TR: 0.65 vs. 0.77 for Appr, 0.58 vs. 0.80 for C&P). This difference is particularly strong for the case of detecting *Bored* states (US vs. TR: 0.67 vs. 0.86 for Appr, 0.59 vs. 0.88 for C&P). Note that we achieved better F1-scores even with the lower number of training samples provided by Turkey experts (672) compared to the

U.S. experts (1776). For *Assessment* models, where we have comparable training set sizes (1574 vs. 1538), although both modalities performed slightly better overall when trained on expert labels from the U.S. compared to Turkey (US vs. TR: 0.53 vs. 0.51 for Appr, 0.74 vs. 0.70 for C&P), we observed that *Confused* state had higher F1-scores when expert labels were obtained from Turkey rather than U.S. (US vs. TR: 0.45 vs. 0.57 for Appr, 0.63 vs. 0.66 for C&P). In Table 4, we observed very similar trends with AUC values as we had with F1-scores for the overall performance results of models trained on expert labels from Turkey versus the U.S.

Table 4. Affect detection classifier results (AUC values) for separate modalities (Appr: *Appearance*, C&P: *Context & Performance*) and different section types (Instr: *Instructional*, Assess: *Assessment*) trained using labels by experts from the United States and Turkey.

Section Type	Labels: Experts from the U.S.				Labels: Experts from Turkey			
	Avg. Train Count	*Total Test Count*	Appr AUC	C&P AUC	*Avg. Train Count*	*Total Test Count*	Appr AUC	C&P AUC
Instr.	*1776*	*536*	0.62	0.57	*672*	*519*	0.66	0.76
Assess.	*1574*	*635*	0.54	0.81	*1538*	*458*	0.53	0.78

4 Discussion And Conclusions

In this study, we explored whether having expert labelers from the same country as the students would lead to more reliable affective state labels than having experts from a different country. Our findings show that experts from Turkey obtained moderately better inter-rater agreement than the experts from the U.S. These results are perhaps unsurprising, given that the original data collected was on learners from Turkey, and experts from Turkey have shared culture (common values and perceptions; modes of expressing affect) and context (similar school experience, shared environment). More importantly, even though the U.S. experts agree with each other, they agree fairly poorly with the Turkey experts, even when only the best U.S. experts are taken into account. Again, this finding may not be surprising. However, these findings have not been previously demonstrated in a quantitative comparison, and cross-cultural affect coding is common in the field of AIED. These results argue that the cross-cultural affect coding should be done with extreme caution.

On the other hand, it should be noted that these two expert groups had slightly different training due to being trained in a different setting and language; the two groups were also somewhat demographically different in terms of their age. This is the type of limitation that is difficult to surmount when comparing experts from different populations, but it suggests that replication with larger samples is probably warranted.

This difference between what the Turkey experts and U.S. experts saw, within the same student data, can also be seen in the differences between the distributions of affective states obtained from the two groups. In particular, these two distinct expert groups interpreted and assigned *Bored* and *Satisfied* states rather differently, with U.S.

experts assigning a label of *Satisfied* much more frequently and *Bored* much less frequently. This may be due in part to the type of affect commonly experienced during online learning. Students' emotions may be less intense in a 1:1 learning using an online content platform (i.e., watching instructional videos or solving assessment questions) than in other contexts [20]. Qualitatively, several experts in both Turkey and the U.S. commented that the students were often very close to the neutral state. To differentiate between student emotional states, the human expert might need to understand subtle signals of shifts in emotion. This issue might explain the differences in the state distributions: There could be a cultural impact in interpreting such ambiguities in close-to-neutral states of students. Alternatively, U.S. experts may simply have been unable to recognize some of the key signs of boredom among the students from Turkey. This possible difference or limitation of U.S. experts in understanding subtle differences between emotions might also explain the higher F1-scores obtained for *Bored* and *Confused* states of students from Turkey when experts also from Turkey are utilized to provide somewhat more reliable ground truth labels for training affect detection models. Although we have limited data in these experiments, considering the challenges of affect detection both for human experts and machines, the differences in model results suggest that it is important to obtain labels as reliable as possible to achieve high-performing detection of student affect. These results might also suggest that although the best three experts from the same country could agree with each other reasonably well, we should be careful when using human experts from a different country than participants as this can impact the final model results.

These findings argue in general, then, that inter-country labeling of student affect is non-ideal. This finding raises some questions for going forward, however. First and foremost is, how different can labelers be from students and still produce acceptable labels? Living in the same country is an easy shorthand for belonging to the same culture, but culture and country do not perfectly coincide. Could a Canadian reliably label American data? Could a New Yorker reliably label a Texan's affect? And how long must someone have lived in the same country as the subject contributing data, to be reliable? What if they are married to a member of the target population? These questions are difficult to answer, but essential if we are to fully leverage this type of finding for data collection for affective computing research. One suggested take-away message from this research is that cross-national or cross-cultural expert labelers should be vetted for inter-rater agreement very carefully (a practice recommended in [11]), but having said that, similar precautions should be taken with any set of experts.

It has been a matter of debate within the field for decades whether it is wise to conduct affect labeling cross-culturally (for example, [9] argues in favor and [11] against). Our conclusion indicates that if a researcher's goal is to obtain high-quality labels of student affect, whether for use in affect detection or analysis on their own, it is probably not ideal to simply take a convenience sample of expert labelers who do not belong to the same population as the research subjects. Ultimately, AIED models and interventions based on those models have the highest chance of being effective if they are based on more reliable data.

References

1. Sabourin, J., Mott, B., Lester, J.C.: Modeling learner affect with theoretically grounded dynamic Bayesian networks. In: Proceedings of International Conference on Affective Computing and Intelligent Interaction, 286-295. Springer, Berlin, Heidelberg (2011).
2. Jaques, N., Conati, C., Harley, J.M., Azevedo, R.: Predicting affect from gaze data during interaction with an intelligent tutoring system. In: Proceedings of International Conference on Intelligent Tutoring Systems, 29-38. Springer, Cham (2014).
3. Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M.: Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. Journal of Learning Analytics, 1(1), 107-128 (2014).
4. Liu, Z., Pataranutaporn, V., Ocumpaugh, J., Baker, R.S.J.d.: Sequences of Frustration and Confusion, and Learning. In: Proceedings of the 6th International Conference on Educational Data Mining, 114-120. International Educational Data Mining Society (2013).
5. D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., Graesser, A.: A Time for Emoting: When Affect-Sensitivity Is and Isn't Effective at Promoting Deep Learning. In: Proceedings of International Conference on Intelligent Tutoring Systems, 245-254. Springer, Berlin, Heidelberg (2010).
6. Arroyo, I., Woolf, B.P., Burleson, W., Muldner, K., Rai, D., Tai, M.: A Multimedia Adaptive Tutoring System for Mathematics that addresses Cognition, Metacognition and Affect. International Journal on Artificial Intelligence in Education, 24 (4), 387-426 (2014).
7. D'Mello, S.K., Graesser, A.C.: Dynamics of Affective States during Complex Learning, Journal of Learning and Instruction, 22 (2), 145-157 (2012).
8. Dillon, J., Ambrose, G.A., Wanigasekara, N., Chetlur, M., Dey, P., Sengupta, B., D'Mello, S.K.: Student affect during learning with a MOOC. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, 528-529. ACM (2016).
9. Ekman, P., Friesen, W.: Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists, Palo Alto (1978).
10. Sayette, M.A., Cohn, J.F., Wertz, J.M., Perrott, M.A., Parrott, D.J.: A psychometric evaluation of the facial action coding system for assessing spontaneous expression. Journal of Nonverbal Behavior, 25 (3), 167-185 (2001).
11. Ocumpaugh, J., Baker, R., Rodrigo, M. M. T.: Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences (2015).
12. Aslan, S., Mete, S. E., Okur, E., Oktay, E., Alyuz, N., Genc, U., Stanhill, D., Arslan Esme, A.: Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement. Journal of Educational Technology, 57(1), 53-59 (2017).
13. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the Facial Action Coding System. The handbook of emotion elicitation and assessment, 203-221. Oxford University Press, New York, NY, US (2007).
14. Stemler, S. E.: A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research and Evaluation, 9(4), 1-19 (2004).
15. Gwet, K.L.: Handbook of Inter-Rater Reliability. Advanced Analytics, LLC (2010).
16. Krippendorff, K.: Computing Krippendorff's alpha-reliability. Departmental Papers (ASC), 43. Retrieved from http://repository.upenn.edu/asc_papers/43 (2011).

17. Siegert, L., Böck, R., Wendemuth, A.: Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. Journal of Multimodal User Interfaces, 8(1), 17-28 (2014).
18. Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., Arslan Esme, A.: Behavioral engagement detection of students in the wild. In: Proceedings of the 18th International Conference on Artificial Intelligence in Education, 250-261. Springer, Cham (2017).
19. Alyuz, N., Okur, E., Genc, U., Aslan, S., Tanriover, C., Arslan Esme, A.: An unobtrusive and multimodal approach for behavioral engagement detection of students. In: Proceedings of the 1st International Workshop on Multimodal Interaction for Education, 26-32. ACM (2017).
20. Lehman, B., Matthews, M., D'Mello, S., Person, N.: What are you feeling? Investigating student affective states during expert human tutoring sessions. In: Proceedings of International Conference on Intelligent Tutoring Systems, 50-59. Springer, Berlin, Heidelberg (2008).