The challenge of gathering self-reported moods: Cases using a classroom orchestration tool

Marc Beardsley, Milica Vujovic, Emily Theophilou, Davinia Hernández-Leo
Department of Information and Communications Technologies
Universitat Pompeu Fabra
Barcelona, Spain
marc.beardsley@upf.edu

Marta Portero Tresserra
Department of Psychobiology and Methodology of Health Science
Universitat Autònoma de Barcelona (UAB)
Barcelona, Spain
marta.portero@uab.cat

Abstract—Self-reports of affective states are increasingly being collected in educational settings. However, individual definitions and usage of emotion and mood terms are often subjective despite objective definitions becoming more widely accepted. We explore whether the variation among individual learners in how mood terms are defined presents an obstacle to using self-reported mood data for group comparison studies. Following a design-based research methodology, we ran two case studies to explore the use of ClassMood App in a multimodal learning study and the validity of the self-reported mood data it collected. During the first case, 24 primary school students experienced difficulty understanding the words used to describe the moods in ClassMood App. In the second case, involving 77 university students, we explored whether the misunderstanding of mood words persisted with older students. Participants were asked to rate their familiarity with and match definitions to a set of 8 mood words. We found that levels of familiarity with the mood words varied greatly and 17.9% of all definition matching attempts were incorrect. The results suggest that the variation in subjective definitions of mood terms is likely to affect the validity of the data collected by the ClassMood App for group comparison studies.

Keywords—Emotion measurement, data validation, social and emotional learning, user interface design, orchestration tool

I. INTRODUCTION

Instruments that incorporate self-reports of affective states are increasingly being used in educational settings [1, 2]. The self-report data collected can potentially be used to assess the effectiveness of educational interventions [3]. However, individual definitions and usage of emotion and mood terms are often subjective despite objective definitions becoming more widely accepted [4]. A difference in how mood terms are defined and applied may affect the validity of self-report mood data when used for group comparison studies – and such studies are commonplace when assessing educational interventions.

Validated self-report instruments for measuring affective states in educational settings include the Medical Emotion Scale (MES) [1] and Achievement Emotion Questionnaire (AEQ) [2]. MES presents participants with words representing affective states and asks them to rate the intensity of the state. AEQ presents statements that include words representing affective states and asks participants to rate their level of agreement with the statements. The words representing affective states are consistent with the circumplex model of affect [5]. It is becoming more widely accepted that objective mood definitions can be represented in the circumplex model of affect. The x-axis of the model is the continuum between emotions considered pleasant and unpleasant, like happiness and sadness respectively. The y-axis is the continuum between high and low levels of arousal. Each of these emotional categories tend to correlate with specific physiological changes in the body [6] and specific neural structures and pathways [7, 8]. In this context, measures of affective state that are independent of individual perception, such as physiological measures, pattern recognition or brain states, are often found to be more accurate and objective than self-reported states [9]. Thus, the variation among individuals in how emotions are defined and applied can affect the validity of self-reported mood data. For example, if my definition of sad is intense while yours is placid, then it is difficult to state that we are referring to the same mood. Nevertheless, self-reports of emotion are useful to the extent that they relate to an individual’s perception of their currently experienced emotions [10] and can contribute to better self-regulated behavior [11].

The importance of improving self-regulated behaviors of learners is evidenced by results of interventions focused on developing students’ social and emotional skills in school education settings. Such interventions have resulted in lasting positive effects on student “skills, positive attitudes, prosocial behavior, and academic performance” [12]. To help teachers integrate social and emotional learning (SEL) activities into their classes, we developed an online, classroom orchestration tool called the ClassMood App [13]. The objective of the ClassMood App is to provide teachers with a tool that can scaffold student development of self-regulatory practices over time. The application focuses on supporting learner development of self-awareness and self-regulatory competencies. These competencies have been identified as priorities in the framework for systemic social and emotional learning proposed by the Collaborative to Advance Social and
Emotional Learning (CASEL) [14]. Developing these competencies involve improving learner proficiency at recognizing one’s emotional state, labelling it accurately (i.e. selecting an appropriate word to describe it), and being able to influence it with self-regulatory practices [15]. These competencies have been selected as understanding emotions strongly relates to wellbeing and academic performance. MacCann et al. (2020) [16] write, “knowledge about the causes and consequences of emotions and a vocabulary of emotions words, along with knowing how to manage emotional situations are potentially the most important parts of Emotional Intelligence for academic performance.” In short, the ClassMood App anonymously collects students’ self-reported moods via a U-shaped graphical interface. As self-reported moods are collected, the tool determines an aggregate class mood and suggests a classroom activity to the teacher in order to influence the mood of the class. Most of the suggested activities model evidence-based self-regulatory practices such as mindful breathing and have been shown to induce changes in the affective states of individuals [17].

Beyond use within the tool to generate an aggregate class mood, the data collected by the ClassMood App is envisioned to be used to assess the effects of the ‘mood-changing’ activities (i.e. self-regulatory activities) suggested by the tool; as an impact measure for intervention programs such as the Spotlighters project (http://spotlighters.eu) that of which the tool is an output of; and as a source of data for multimodal learning research. Moreover, identifiable data could be collected and used in the future to track changes to student moods over time. This richer data could more accurately measure the effectiveness of the learning activities and interventions based on student profiles; and possibly be used as multimodal training data (e.g. combined with physiological measures, auditory and visual modalities) for emotion recognition systems. For these purposes the validity of the data is crucial. Previous studies have validated the use of self-report instruments for measuring affective states in educational settings [1, 2]. Yet these studies did not test participant understanding of the words used to describe the affective states. Our study, which is framed within a design-based research iterative cycle focused on improving interface designs for gathering self-reported moods, provides an investigation into whether the variation among individual learners in how mood terms are defined presents an obstacle to using the self-reported mood data collected by the ClassMood App for group comparison studies. In the section that follows, we present two case studies interceded by a description of the iterative improvements made to the ClassMood App interface.

II. Approach

As the goal of the ClassMood App is to make it applicable to multiple educational contexts from primary school to higher education; and available in multiple languages, we follow a multiple case study approach [18]. The cases are framed in a design-based research methodology [19], which allows iterative feedback cycles of developing and testing increasingly satisfying solutions to problems. As part of this approach, we report two case studies and focus our analyses on the ClassMood mood measure interface that is used to collect the self-reported moods of participants (see Fig. 1).

The main purpose of the mood measure interface is to collect actionable data for the educator rather than accurately represent the spectrum of moods a student may be experiencing in class. In other words, the goal is to have students pick a positive or negative valence and high or low arousal mood rather than a neutral mood to better facilitate an activity suggestion.

Consequently, the initial design of the interface is limited to 8 moods to limit the choices presented to students (see Fig. 1). The U-shaped graphical interface combines concepts of the Yerkes-Dodson Law [20, 21] and affective circumplex model [5]. The affective circumplex model has been used to inform other instruments and applications that incorporate self-report mood interfaces [1, 2, 22]. The final selection of mood words was made by collaborating neuroscientists who selected the words that most accurately represented their placement on the U-shaped interface. Further, the words were translated into Catalan by a neuroscientist to match the local context in which the studies were to be run. Finally, the colors used in the interface were derived from research that suggests the link between emotions and colors is rooted in biology [23].

A. Case 1: Elementary School Students

In the first case, four separate groups of six school children (n = 24; 4 male and 20 female; Ages 6 to 9) participated in an observational study exploring the collection and interpretation of multimodal learning data. The study, conducted in Catalan, involved the collection of physiological data (e.g. galvanic skin response); video, audio and motion capture recordings; qualitative data from researcher and instructor observations; and ClassMood data. The ClassMood data was mainly being explored for its suitability as a before-and-after measure for detecting the effects of the mood changing activities being considered for integration into the tool.

During the first case, the young participants experienced difficulty in understanding the mood measure interface the first time they were asked to use the tool. Specifically, they were unfamiliar with some of the words used to describe the moods and required an explanation of what the words meant.
Documented comments by observers (i.e. researchers and instructors) can be summarized as follows: students found the interface to be easy to use but found it challenging to understand the words used to define some expressions, especially Ansios (Anxious) and Frustrat (Frustrated). Further, the words Alegre (Joyful) and Entusiasmat (Excited) were too similar to be distinguishable.

B. Updates to the ClassMood Mood Measure Interface

After the first case study experience, efforts were made to improve the mood measure interface. The selected mood words were revised in an attempt to better facilitate the capturing of the category of the mood or core affect [24] rather than trying to accurately capture a more narrow and specific mood. To facilitate the capturing of core affect, the graphical interface design changed from a U-shape representation to a grid structure (Fig. 2) with the x-axis representing valence and y-axis representing arousal. Half of the moods are of positive valence and half of negative valence. Further, half of the moods are positive arousal (i.e. high energy moods) and half are negative arousal (i.e. low energy moods). The selected words feature antonyms along the x-axis to make it easier to distinguish the moods and to better provide a context for understanding the definitions of the mood words.

C. Case 2: First-Year Undergraduate Students

In the second case study, we set out to explore whether the problem of understanding mood words persisted with university students. To do so we conducted a survey study involving first-year undergraduate students (n = 77; 21 female and 56 male) at a university in Barcelona, Spain. Participants were presented Catalan translations of the updated mood words in an online survey form and asked to rate their level of familiarity with the words; and then match definitions to them. Participants were able to select multiple definitions for each word. The definitions were derived from theories on Core Affect [21]. The evaluation of the definition responses was based on Table I. The table presents the exact match for a word and definition; and the close matches. The close matches are the words that have similar but not identical definitions. The words that do not appear as exact or close matches were deemed to be incorrect. A 5-point Likert scale was used to collect ratings of familiarity.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Exact Match</th>
<th>Close Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Having lots of energy with unpleasant feelings.</td>
<td>Agitated</td>
<td>Upset</td>
</tr>
<tr>
<td>Having lots of energy with pleasant feelings.</td>
<td>Excited</td>
<td>Happy</td>
</tr>
<tr>
<td>Having energy with unpleasant feelings.</td>
<td>Upset</td>
<td>Agitated, Sad</td>
</tr>
<tr>
<td>Having energy with pleasant feelings.</td>
<td>Happy</td>
<td>Excited, Content</td>
</tr>
<tr>
<td>Having little energy with unpleasant feelings.</td>
<td>Sad</td>
<td>Upset, Tired</td>
</tr>
<tr>
<td>Having little energy with pleasant feelings.</td>
<td>Content</td>
<td>Happy, Sleepy</td>
</tr>
<tr>
<td>Having no energy with unpleasant feelings.</td>
<td>Tired</td>
<td>Sad, Sleepy</td>
</tr>
<tr>
<td>Having no energy with pleasant feelings.</td>
<td>Sleepy</td>
<td>Content, Sad, Tired</td>
</tr>
</tbody>
</table>

III. RESULTS

Upon completing the study, it was determined that ‘contingut’ was a mistranslation of the English word ‘content’ and the subsequent analysis is of the remaining 7 words.

A. Analyses of words

To assess the familiarity of each mood word, we calculated the average rating per word and, also, determined the number of participants that selected 4 or 5 on the 5-point familiarity scale. The results are shown in Table II. The average rate of familiarity across all words was 72.3% and ranged from a low of 44.2% for Agitat (Agitated) to a high of 93.5% for Cansat (Tired).

To assess how well the mood words were defined, we determined whether the matching definitions selected by participants were exact or incorrect matches (see Table I). For the exact match analysis, we assigned 1 point for each correct match. For instances in which participants had a correct match but had selected multiple definitions for a single mood, their 1 point was divided by the number of responses given (e.g. 1 correct and 2 incorrect answers resulted in a score of 0.33). The results are presented in Table II and show that across the 7 mood words, an exact match of the word to its definition occurred in 51.37% of the responses with the highest rate of exact matches being 74.7% for Emocionat (Excited) and lowest being 31.9% for Cansat (Tired). To determine the incorrect matches, 1 point was assigned for exact or close matches (see Table I) and 0 points for the remaining answers deemed to be incorrect. Overall 18.9% of responses were incorrect with the highest rates of incorrect answers occurring in words representing unpleasant states (e.g. Trist, Agitat, Cansat).

B. Analyses of participants

The scoring approach described above was maintained for the per participant analyses. The rates of familiarity by participant can be found in Table III and range from a low of being familiar with 1 out of 7 mood words (1.3% of participants) to a high of being familiar with all 7 mood words (31.2% of
participants).

### Table II. Per Word Familiarity Ratings

<table>
<thead>
<tr>
<th>Word (English)</th>
<th>Familiarity</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalan</td>
<td>Average Rating</td>
<td>Familiar with</td>
</tr>
<tr>
<td>Agit (Agitated)</td>
<td>3.21</td>
<td>44.2%</td>
</tr>
<tr>
<td>Emocionat</td>
<td>3.94</td>
<td>74.0%</td>
</tr>
<tr>
<td>Molest</td>
<td>3.66</td>
<td>58.4%</td>
</tr>
<tr>
<td>Content</td>
<td>4.22</td>
<td>85.7%</td>
</tr>
<tr>
<td>Trist (Sad)</td>
<td>3.66</td>
<td>62.3%</td>
</tr>
<tr>
<td>Contingut</td>
<td>3.18</td>
<td>37.7%</td>
</tr>
<tr>
<td>Cansat (Tired)</td>
<td>4.38</td>
<td>93.5%</td>
</tr>
<tr>
<td>Dormit (Sleepy)</td>
<td>4.30</td>
<td>88.3%</td>
</tr>
</tbody>
</table>

C. Familiarity and Correctness

A Spearman's rank-order correlation was run to determine the relationship between familiarity ratings and exact matching definitions ($r(75) = -.32, p = .48$) and incorrect definitions ($r(75) = -.04, p = .94$). No significant correlations were found (see Fig. 3).

### Table III. Aggregation of Participant Familiarity Ratings and Definition Scores

<table>
<thead>
<tr>
<th>Word count</th>
<th>Familiarity</th>
<th>Exact Matches</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>31.17%</td>
<td>9.09%</td>
<td>0.00%</td>
</tr>
<tr>
<td>6</td>
<td>20.78%</td>
<td>5.19%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>11.69%</td>
<td>15.58%</td>
<td>2.60%</td>
</tr>
<tr>
<td>4</td>
<td>9.09%</td>
<td>16.88%</td>
<td>3.90%</td>
</tr>
<tr>
<td>3</td>
<td>15.58%</td>
<td>22.08%</td>
<td>7.79%</td>
</tr>
<tr>
<td>2</td>
<td>10.39%</td>
<td>19.48%</td>
<td>22.08%</td>
</tr>
<tr>
<td>1</td>
<td>1.30%</td>
<td>9.09%</td>
<td>41.56%</td>
</tr>
<tr>
<td>0</td>
<td>0.00%</td>
<td>2.60%</td>
<td>22.08%</td>
</tr>
</tbody>
</table>

Further, 40 (52%) of participants rated being familiar or highly familiar with at least 6 of the 7 words; while 21 (27.3%) rated being familiar or highly familiar with less than half of the words. The average across all participants was a rating of being familiar or highly familiar with 3.9 of the 7 words.

Seven participants (9.1%) were able to exactly match all words to their definitions; 23 (29.9%) matched at least 6 words; while 41 (53.2%) matched less than half of the words with their exact definitions. The average across participants was exactly matching 3.6 of the 7 words with their definitions. In looking at the rate of incorrect answers given, 17 (22.1%) gave no incorrect answers; 66 (85.7%) gave two or fewer incorrect answers; while 5 (6.5%) gave 4 or more incorrect answers. The average across participants was a rate of 1.3 incorrect definition matches out of 7 possibilities.

A. Limitations

The main limitation of the study is its generalizability. The study was conducted in Catalan. The participants in the survey study were first-year university students in an engineering degree program. It is important to explore whether similar results emerge in studies conducted in other languages and contexts.

In looking at the rate of incorrect answers given among participants from a high of 34.6% for Agit (Agitated) and a low of 7.8% for Emocionat (Excited). Over all, the rate of errors would affect the usefulness in terms of accuracy and comparability of the data collected. Further, on average, participants rated being familiar or highly familiar with only 3.9 out of 7 words; and 28 participants (36.4%) were familiar with 4 or fewer words. Low ratings of familiarity suggest that we did not select or translate words correctly. For example, Contingut (Content) was mistranslated and resulted in the majority of participants (66.3%) incorrectly defining the word. In general, the results suggest that the problem of misunderstanding words to describe affective states, brought to light in the first case study with elementary school children, persisted with university students. Also, there was a wide range in the rate of incorrect responses given among participants from a high of 5 out of 7 (2 participants) to a low of 0 (17 participants). This suggests that there may not be a straightforward solution and supports the notion that there is a need to improve student understanding and recognition of affective states (i.e. self-awareness).
II. CONCLUSION

We explored the possibility of using the student self-report mood data collected by the ClassMood App for multimodal learning research. The application could serve as both a practical tool for educators and as a validated instrument for research. The results of our exploratory study suggest that the variation in subjective definitions of mood words by participants is likely to affect the validity of the data collected by the ClassMood tool. For example, if one participant defines a mood word much differently than another, then it cannot be assumed that these participants are referring to the same construct. Further work is needed to improve the accuracy of the self-reported moods collected via the ClassMood App.

III. FUTURE WORK

As instruments that incorporate self-reporting of affective states are increasingly being used in educational settings, there is a need to operationalize the variation in how words are used to describe affective states in such instruments and explore approaches that allow self-reported states to be calibrated to the individual to facilitate more accurate comparisons within and between groups. Further, it would be beneficial to explore factors that may influence the discrepancy in defining affective states such as language, demographics and context. Efforts to improve the validity of the self-report mood data collected via the ClassMood App could be pursued through complementary research and student training related to self-awareness (i.e. recognition and labelling of affective states); interface design; or even through the creation of an application that can facilitate the assessment and comparison of items used for self-reported mood interfaces be they words, emoticons, or colors.

ACKNOWLEDGMENT

This work has been partially funded by the European Regional Development Fund, Erasmus+, and the National Research Agency of the Spanish Ministry of Science, Innovation and Universities, under project grants TIN2017-85179-C3-3-R, 2018-1-ES01-KA201-050646, and MDM-2015-0502. D. Hernández-Leo acknowledges the support by ICREA under the ICREA Academia programme. We would like to thank the Spotligheters project team for their contributions to the ClassMood App.

REFERENCES