Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol 6 No. 3 Special issue (2018) 37- 56
ISSN 2295-3159

Do we betray errors beforehand? The use of eye tracking, automated face recognition and computer algorithms to analyse learning from errors

Christian Harteis[a], Christoph Fischer[a] , Torben Töniges[b], Britta Wrede[b]

[a] Paderborn University, Germany
[b]Technical Faculty, Bielefeld University, Germany.

Article received 14 May 2018 /revised 25 September /accepted 26 September /available online 7 December

Abstract

Preventing humans from committing errors is a crucial aspect of man-machine interaction and systems of computer assistance. It is a basic implication that those systems need to recognise errors before they occur. This paper reports an exploratory study that utilises eye-tracking technology and automated face recognition in order to analyse test persons' emotional reactions and cognitive load during a computer game and learning through trial and error. Computer algorithms based on machine learning and big data were tested that identify particular patterns of test persons' gaze behaviour and facial expressions that antecede errors in a computer game. The results show that emotions and learning from errors are positively correlated and that gaze behaviour and facial expressions inform about the errors that follow. However, the algorithms still need to be improved through further studies to be suitable for daily use. This research is innovative in its use of mathematical formulae to operationalise learning through errors and the use of computer algorithms to predict errors in human behaviour in trial-and-error situations.

Keywords: face recognition; eye tracking; emotions; learning from errors

Corresponding author: Prof. Dr. Christian Harteis Warburger Str. 100, 33098 Paderborn christian.harteis@upb.de
DOI: https://doi.org/10.14786/flr.v6i3.370

1. Introduction: Research problem

Working life becomes increasingly complex and challenging, particularly through technological development and digitalisation that aim to enable flexible work processes. As a result, the organisation of work is changing, as well as – in consequence – working tasks and working tools. These become more difficult and the need for efficacy may generate time pressures. Under these conditions, the risk of errors arises. Estimations differ of the amount of worktime spent on errors in enterprises but are as high as half of the entire worktime (Hofman & Frese, 2011). On the one hand, systems engineering can strive to develop intelligent systems that prevent humans from committing errors. On the other hand, research into complex systems has revealed that it is not possible for human activity to avoid errors completely. Hence, learning from errors becomes a relevant issue because it is at least possible to avoid the repetition of errors (Harteis & Bauer, 2014). There is no contradiction in simultaneously trying to develop systems that prevent errors (as far as possible) and postulating learning from errors, because both issues are interrelated. Rather, understanding how to learn from errors is a precondition for developing man-machine interaction that assists in error avoidance. Hence, the main research problem addressed here is how to understand learning from errors in order to provide safety through error prevention in man-machine interaction.

Theoretically, learning from errors has the following preconditions, none of which is trivial in the context of work (Bauer & Mulder, 2013; Oser & Spychiger, 2005): (a) the error has to be identified, (b) feedback to the acting person has to occur and (c) reflection and cause analyses have to result in the creation of negative knowledge – that is, knowledge about how things are not shaped and how processes do not work (Gartmeier, Bauer, Gruber, & Heid, 2008; Oser, Naepflin, Hofer, & Aerni, 2012). In addition, during these processes, the individual concernment of the failing person has to occur. "Concernment refers to the emotional reaction, in which the error embarrasses the actor in a certain way. Such an emotional reaction adds value to the experience of the error situation" (Harteis & Bauer, 2014, p. 710). This added value attaches sufficient importance to the error to initiate the cause analysis – subjectively unimportant events can easily be neglected – and adds authority to the knowledge resulting from reflecting on the cause analyses, which ultimately supports appropriate storage in the memory by amplifying the episodic memory and, thus, learning from error that prevents from its repetition (Oser et al., 2012). However, while knowing that concernment resulting in emotional engagement is a crucial precondition for learning from errors, there is so far no evidence about the kind of valence of emotions (e.g. positive or negative) that best supports learning from errors. To sum up: Any kind of emotional reaction in an error situation can be considered a basic precondition for learning from errors.

It was Oser who identified situations that almost – but not finally – ended up with errors as incidents of interest for their suggestion of a sense of failure (Oser, Müller, Obex, Volery, & Shavelson, 2018; Oser & Obex, 2015). Of course, in order to prevent errors, it is important not only to learn from errors but also from incidents in which errors nearly occur. In general, the crucial moment of learning from near misses is the emotional reaction that arises when somebody realises that an error is about to occur or that an error almost happened. Cause analysis and reflection upon the incident play a similar role here as they do for learning from errors. However, investigations into this sense of failure revealed that emotional reactions that accompany (almost) error situations do not necessarily occur as a reaction to the incident itself but may arise shortly before the error occurs (Oser & Volery, 2012). Whereas emotional reactions after an error tend toward embarrassment, cognitive load is considered to be the reason for an emotional reaction before an (almost) error situation (De Jong, 2010). Cognitive load refers to the working load within the limited capacity of the short-term memory (Sweller, 1994). When cognitive load becomes too big, the actor's capacity for information processing and problem perception decreases so that errors become probable.

To investigate the research problem stated above, several issues have to be considered: Emotional reactions and cognitive load are important phenomena in relation to the occurrence of errors or near misses and are important indicators that can help to prevent upcoming errors. The challenge, of course, lies in how best to operationalise and measure these phenomena.

Empirical research on learning from errors and near misses has to date applied self-reporting methods, that is, interviews and questionnaires. Investigating error situations in work contexts is particularly challenging because companies tend to avoid publishing business processes. There are studies investigating employees' attitudes towards errors at work (e.g. Hetzner, Gartmeier, Heid, & Gruber, 2011) which make use of standard self-report questionnaires (e.g. the Error Orientation Questionnaire – Rybowiak, Garst, Frese, & Batinic, 1999) and there are studies investigating ways of dealing with error situations in daily working life which apply self-report questionnaires or interviews (e.g. Harteis, Bauer, & Gruber, 2008). The current state of research thus has to acknowledge the following problems:

- Studies operating with questionnaires focus either on general attitudes towards errors or they introduce a constructed error situation (e.g. through case stories or vignettes) and ask for potential reactions. Neither option provides any insight into how a person actually behaves and reacts in a real error situation. In addition, whether the constructed situation causes a similar emotional engagement to real situations remains a matter for speculation.
- Studies focusing on incidents that participants actually experienced usually ask test persons for episodes in which an error occurred and ask them to describe how the people concerned dealt with this situation. However, it is difficult to relate different cases described by different test persons to each other because the cases themselves represent error situations of different dimensions, because it is unclear how representative the described cases are for the test persons' (work) environment and because the descriptions are probably subjectively biased.
- By their nature, self-reports feature only those mental and emotional processes that test persons are aware of and can remember. Those studies therefore neglect whatever may remain unconscious or cannot be recalled.

Obviously, there is a research gap in the studies that investigate learning from errors, requiring a study that (a) on the one hand reliably provides stable and repeatable error situations for an entire sample, (b) does not depend on subjective biases and memory performances, and (c) is able to grasp unconscious emotional reactions. This study aims to test particular online measures of emotional reactions and cognitive load.

2. Research questions and study context

In order to reach the research aims, research questions were formulated that address theoretical issues and issues of online measurement.

Since field studies have to accept the problems described above, a laboratory setting appears appropriate to establish conditions that are identical for all test persons and exactly repeatable. Admittedly, a laboratory environment lacks authenticity. However, it should be acceptable as long as test persons develop concernment when failing during the experiment. Hence, the present study used a regular jump-and-run computer game (*Give Up 2* by Armor Games) and controlled for test persons' involvement.

2.1 Research questions

The research questions for this study can be separated into thematic (RQ 1 and RQ 2) and methodological (RQ 3 and RQ 4) questions.

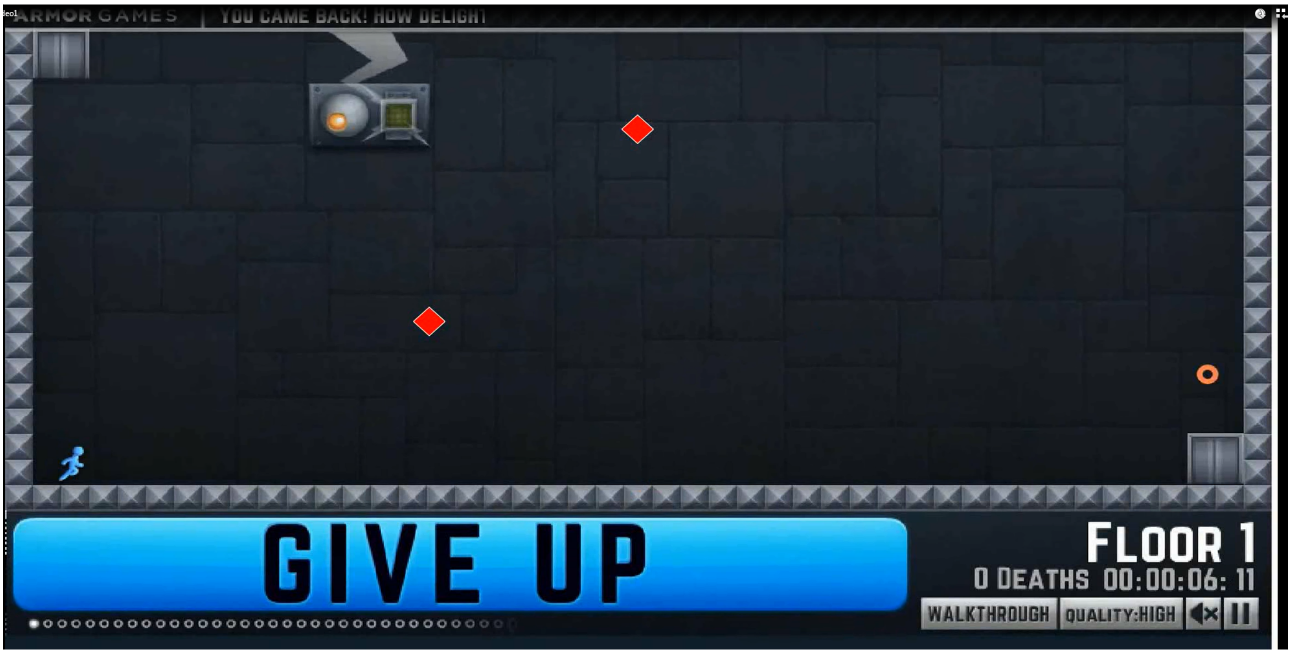
- *RQ 1*. Are there emotional reactions and indicators for cognitive load to be found that precede errors? The hypothesis to be tested is that emotional reactions and/or cognitive load precede errors. An answer to this question is relevant for the intention to anticipate errors before their occurrence.
- *RQ 2*. Is the quality of learning related to emotional reactions? The hypothesis to be tested is that better learners show stronger emotional reactions than worse learners do. This question tests Oser's theory about the importance of concernment for learning from errors.

- *RQ 3*. Are there appropriate online measures that indicate emotional reactions? This study also aims to test particular online measures for their suitability for educational research questions.
- *RQ 4*. Is there a specific measure for the quality of learning from errors? A computer game with exactly repeatable conditions allows the combination of indicators for learning from errors and degree of difficulty.

2.2 Description of the computer game

Give Up 2 is a computer game in which the player operates a figure that needs to overcome obstacles and dangers on various levels of increasing difficulty. Two kinds of error can occur in this game: (a) the player fails to overcome the obstacle with the figure or (b) the figure gets attacked by weapons and dies. In that case, the player has to start from the beginning of the respective level again. The course of action always remains the same, i.e. each player faces the same conditions.

Video 1. Demonstration of the computer game



- Video link: https://www.youtube.com/watch?v=xoOe5Lh1aZw -

This game permits the introduction of different test persons into comparable problem settings that provoke errors. It provides a competitive scenario that should motivate the test persons to perform as well as possible.

3. Methods of data collection, analyses and challenges

This main section of this paper describes the procedures of data collection, data preparation and data analysis. First, a flowchart illustrates the sequence of data processing. Then, the system configuration and the sample description provide insight into the way the data collection was realised (3.1). The raw data were then prepared (3.2) for further analyses, exploring the predictors of errors (3.3) and learning from errors (3.4). The description of the analyses applied here also comprises a discussion of the challenges, because novel approaches were tested. Figure 1 shows a flowchart of the procedures that were applied for this investigation.

They comprise regular and well-established approaches to eye and face analysis utilising particular parameters and procedures that will be explained within this section. The flowchart presents the sequence showing how the data were prepared and analysed.

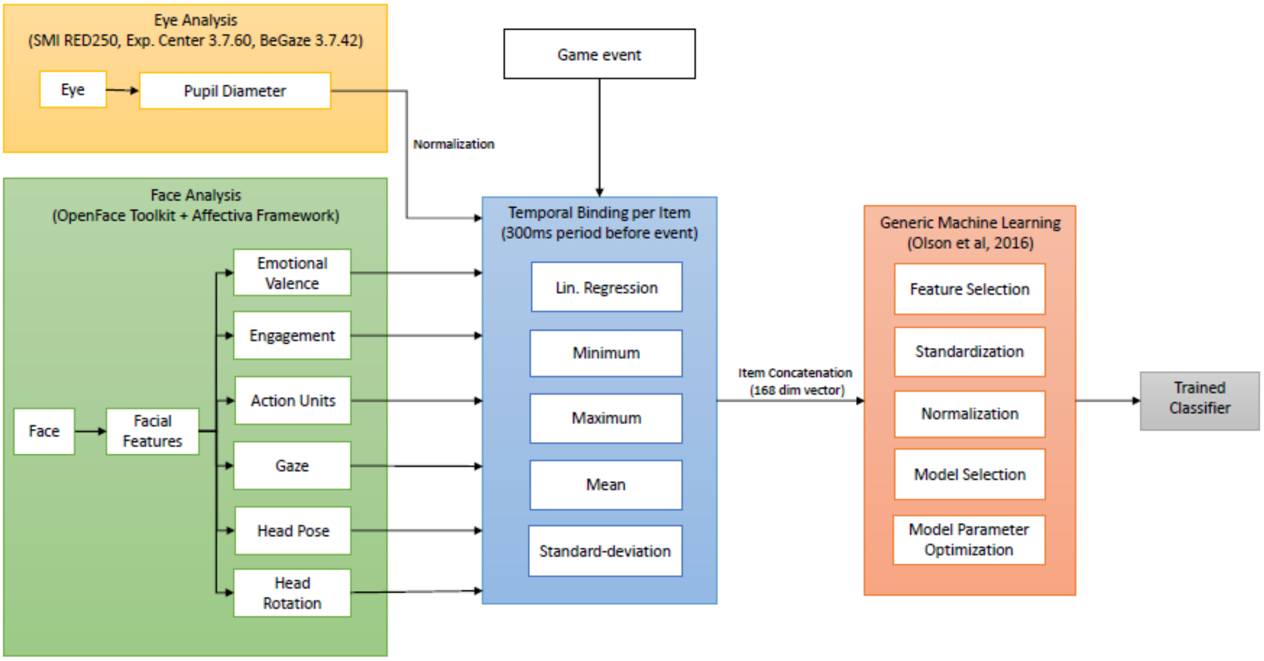


Figure 1. Flow chart of data acquisition and processing.

For eye analyses, pupil diameters were used to grasp cognitive load, and face analyses used standard tools (i.e. the OpenFace toolkit and Affectiva framework) that apply particular parameters to identify emotional reactions. These data were aggregated to temporal bindings per item, which were then used for machine learning. The result of the data-mining process is a trained classifier for the prediction of an error.

3.1 Test procedure

Data collection was realised with the remote eye-tracking system SMI RED250, using the software versions Experiment Center 3.7.60 and BeGaze 3.7.42 and a video camera (Logitech C922 Pro Stream) installed at the top of the stimuli screen within a laboratory, with a headrest and steady artificial lightning, i.e. robust laboratory conditions with no brightness differences between test persons (Holmqvist et al., 2011). Additionally, the stimulus itself did not vary a lot in luminesce. Thirty-eight test persons with varying experience in computer gaming voluntarily took part in the experiment.

Table 1

Sample description.

Category	**Description**
Number of test persons	38 (23 female, 15 male)
Age	Mean: 29.79 (SD: 5.92); min: 20, max: 45
Experience: When was the last time you played a computer game?	Less than a year ago: 25 More than a year ago: 13
Involvement: 7 point Likert scale with 10 items: 7 = high involvement, 1 = low involvement	Mean: 4.50 (SD: 0.92); min: 2.1, max: 6.5

Before starting the experiment, the test persons filled in the consent form, received a video introduction to the game and were asked to confirm that they understood the game and the operation of the system. They also completed a questionnaire before and after the experiment describing their gaming experience and their engagement with the game measured by the Personal Involvement Inventory (Zaichkowsky, 1994). Table 1 indicates that the test persons were sufficiently engaged. The test persons played the game for five minutes on the stimuli presentation screen using three keys of a keyboard, and they were observed by a video camera and remote eye-tracking system. The following in situ-data were generated and utilised:

- *Game video*. Via screen recording, a video of the game was generated, including all inputs of the test persons during the game.
- *Facial video*. The video camera recorded the test person's face during the game.
- *Pupil diameter*. The remote eye-tracking system constantly recorded the test person's right pupil diameter during the game. Changes in pupil diameter apply as indicators of cognitive load (Szulewski, Kelton, & Howes, 2017).

3.2 Data preparation

The synchronisation of these data was realised by time stamps implemented through the eye-tracking software. The challenge for subsequent analyses was to derive meaningful information from these data. Therefore, they were further edited. As a first step, comprehensive annotations were added to the game video:

- *Errors.* Every event in which a test person failed (i.e. being hit by an object or failing to overcome an obstacle) was marked as an 'error'.
- *Successes.* Every event in which a test person succeeded in avoiding an object or overcoming an obstacle was marked as a 'success'.

3.3 Analyses exploring predictors of errors (emotional reactions and cognitive load)

The face videos were analysed by applying the Affectiva framework (McDuff, El Kaliouby, & Picard, 2015) and the OpenFace toolkit (Baltrušaitis, Robinson, & Morency, 2016). Based on millions of facial recording and facial images, these frameworks are able to extract crucial facial landmarks, such as eyebrows or mouth contours (see left side of Video 2). These extracted points on each image are used to extract the head pose, gaze and facial action units (AUs). The Facial Action Coding System (FACS, Ekman & Friesen, 1978) is a taxonomy for classifying various facial behaviours (e.g. AU1: Inner brow raise), and the frameworks used here were able to classify up to 17 of these AUs (Wolf, 2015). The Affectiva framework also combines different AUs to build up expression classes that are more abstract, such as emotional valence (i.e. the positive or negative nature of an emotion) and engagement (the expressiveness of the emotions). These frameworks, particularly in combination, provide derived data on crucial landmarks of the face representing emotional reactions as well as their valence and engagement.

The next step of analysis aimed at identifying the precursors of errors. Data-mining procedures were applied that utilised the derived data from the emotional reactions and a binary classification was modelled. All recorded videos were divided into snippets of 300 milliseconds in length with an overlap of 150 milliseconds. The positive class within the classification was modelled as 'error predication' and all snippets occurring before errors were marked. The negative class was modelled as 'all the rest' and all other data were assigned accordingly. To avoid noisy data, the negative class was filtered by removing those video snippets that followed an error event. The data were split into training (75%) and testing (25%) sets, each set containing the same percentage of positive and negative data.

For each frame of the 300-millisecond snippets, the following 28 items were extracted by OpenFace and Affectiva:

- *17 AUs* - i.e. facial expressions.
- *Valence*.
- *Engagement.*
- *Gaze angle X + gaze angle Y.* Extracted in radians and averaged for both eyes. A person looking from the left to the right results in a change of gaze of angle X, while a person looking from up to down results in a change of gaze of angle Y. If a person is looking straight ahead, both angles will be close to 0 (see Figure 2).
- Normalised *pupil diameter*. If the surrounding lighting conditions are steady, the pupil diameter (see Figure 3) can be used as an indicator of cognitive workload: the wider the diameter, the higher the workload of the person (Beatty, 1982; Krejtz, Duchowski, Niedzielska, Biele, & Krejtz, 2018; Laeng, Sirois, & Gredebäck, 2012; Szulewski, Kelton, & Howe, 2017).
- Head pose *Tx*, *Ty*, *Tz* (location of the head).
- Head rotation *Rx* (pitch), *Ry* (yaw), *Rz* (roll) – see Figure 4.

Figure 2. Gaze angles.

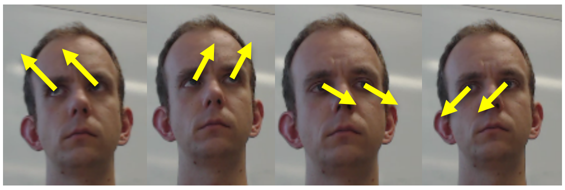


Figure 3. Pupil diameter.

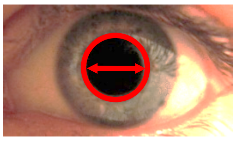


Figure 4. Head rotation.

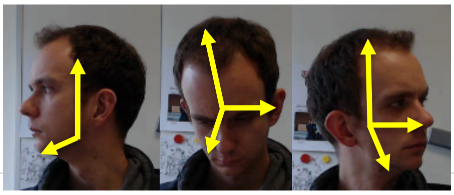
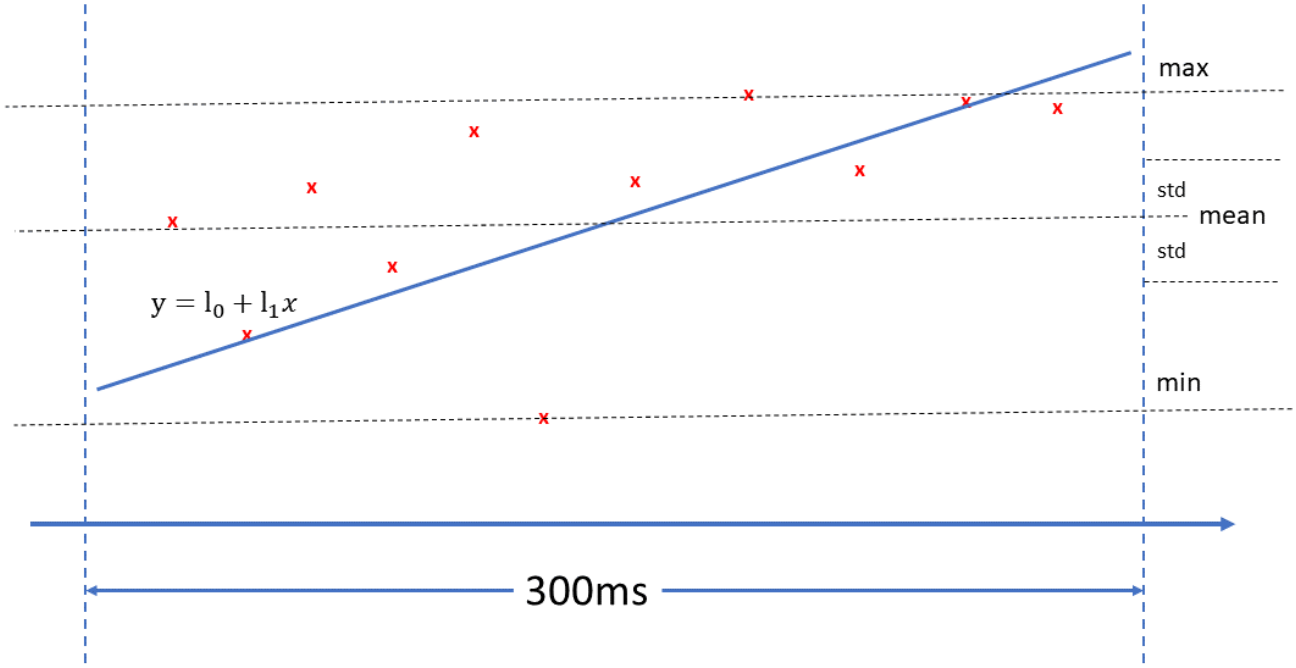


Figure 5. Example of a linear regression for items within 300ms frame.

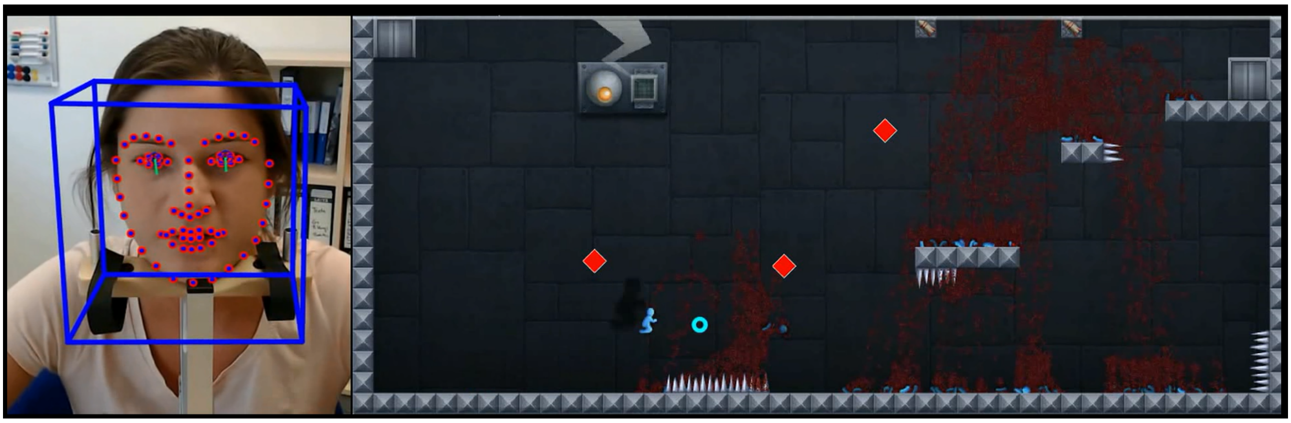


To take temporal dynamics into account, these items were further processed. For each of the 28 items, the temporal and stochastic variations were extracted. Figure 5 shows an exemplary schematic representation: The red crosses represent items occurring within the respective timeframe. For each item, the following features were extracted:

- Linear regression parameters (l_0, l_1)
- Maximum value (max)
- Minimum value (min)
- Mean
- Standard deviation (std)

This results in 6 features per item, totalling 168 features, which were used as inputs for the machine training procedure. Genetic programming (Olson, Urbanowicz, Andrews, Lavender, Kidd & Moore, 2016) was used to train an optimised machine-learning pipeline on the training set that would subsequently be used to evaluate the testing set. The pipeline consisted of multiple steps, such as feature selection, preprocessing, model selection and parameter optimisation. Ultimately, the learned classifier could be used to classify unknown video snippets with the aim of identifying the precursors of errors. While this description of procedures describes the general ratio of analyses, combinations and sets of features were also varied in order to explore answers for the research questions raised above.

Video 2. Demonstration of aggregated data.

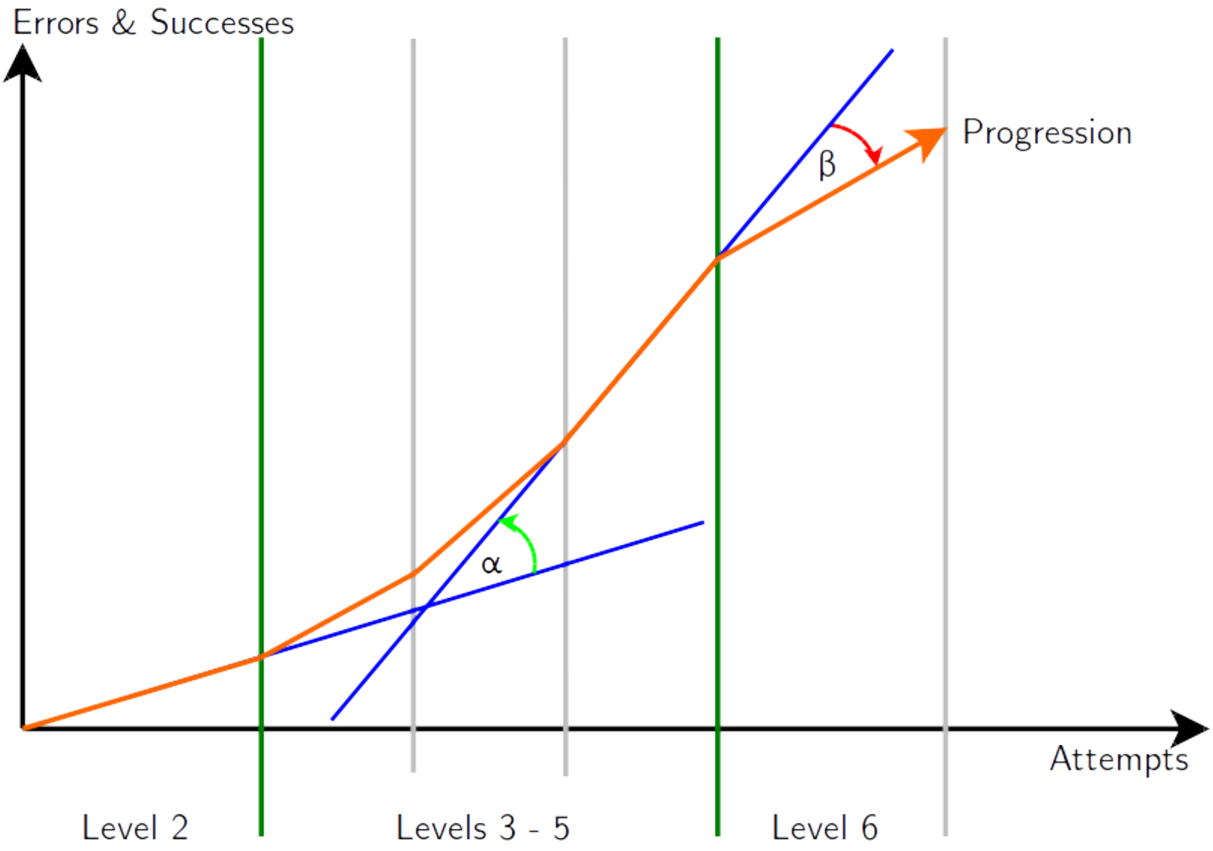


- https://www.youtube.com/watch?v=TTT5sM-m4eM&t=2s -

3.4 Analyses exploring learning from errors

At this point, two kinds of derived data had been generated: errors and successes on the one hand, and emotional reactions on the other. As a third kind of derived data, the quality of learning from errors had to be identified. Disciplines that investigate learning through a formalised lens have developed the idea of calculating a learning curve to indicate quality of learning (Jaber, 2016; Yelle, 1979). Applied to the computer game, the learning curve (orange) can be defined by indicating the errors and successes (y-axis) of a specific task for all attempts (x-axis) at solving the task on the different game levels (separated by vertical lines; see Figure 6).

Figure 6. Example of a learning curve.



As a computer game requires a quite specific kind of learning of one particular skill (i.e. mastering the task), a formalised perspective for differentiating qualities of learning appears appropriate. An overall comparison of the ratio of successes and errors would provide a measurement of the overall performance of the player but would grant no insight into skill development. In order to grasp this, it is necessary to evaluate the performance progression over time. However, changes in performance indicated by changes in the slope of the learning curve can be illustrated by an angle (in Figure 6: α and β). These angles indicate improvement if they have a positive value and a decline if they have a negative value. The design of the computer game provides a remarkable increase in playing difficulty at levels 2 and 6. Each time the difficulty increases, the player has to readjust their behaviour and thereby improve their skill. Hence, at levels 2 and 6, we can expect relatively more errors, compared to successes, than at levels 3, 4 and 5. The introduction of a new task problem provides the player with an opportunity to learn or improve their skill. The learning curve is thus steeper at levels 3, 4 and 5 than at levels 2 and 6. A possible approach to calculate learning quality contrasts the angles of the learning curve between levels 2 and 5 and between levels 5 and 6. Level 5 represents the last easy playing level; the learning curve is considered to be steepest here. Levels 2 and 6 represent difficult playing levels. Looking at the absolute slope of the learning curve would be biased in favour of test persons who started with an already high skill level; however, it is not the intention to measure a test person's absolute skill level but their skill development. Hence, the slope at level 2 defines the baseline skill the test person shows after the first increase in difficulty. The following three levels of a similar difficulty provide the test persons with opportunities to improve on the lower level of difficulty. The difference between the slopes at levels 2 and 5 (angle α) represents the skill increase (or decrease) during this phase of the game. However, focusing on this difference alone would be biased in favour of test persons who performed very weakly at level 2 but who improved at level 5. As the change in skill is also relevant for difficult tasks, it is necessary to consider the development during level 6 – represented in the difference between the slope at levels 5 and 6 (angle β). A consideration of both angles corrects the bias of angle α towards a weak performance at level 2 combined with a good performance at level 5. Hence, the derived data on learning quality considers the relative increase in successes between levels 2 and 5 in contrast to the relative increase in errors between levels 5 and 6. To make this measurable, the two angles α and β are calculated, where α is the angle between the slopes of levels 2 and 5 and β is the angle between the slopes of levels 5 and 6. The quality of learning can thus be defined by the following formula:

$\mathit{Learning} = \alpha + \beta$

Hence, at this step of data preparation, the following derived data was available in order to answer the research questions:

- Performance: *Errors* and *successes*
- Emotional reactions: *Valence* and *engagement*
- Cognitive load: changes in pupil diameter
- Quality of learning: *Learning* = α + β

The methodological challenges were to overcome the weaknesses of previous research on learning from errors as discussed in the section above. One of the major concerns raised there was that previous research does not inform about factual learning from errors. Of course, a computer game provides quite a specific scenario for learning from errors. However, the major advantage is that it provides stable and repeatable conditions for all test persons. The gaming situation provides an opportunity to establish experimental conditions that appropriately motivate test persons to perform as well as possible while allowing them to commit errors without serious consequences. Hence, the computer game provides an experimental scenario in which actual reactions to error can be observed.

There is a further concern about the reliability and validity of data on learning from errors. In this experiment, particular online measurements were implemented to indicate learning from errors. It is a challenge, of course, to derive meaningful data from the data which are themselves extracts from raw data. The quality of these derived data will be discussed in later sections.

4. Results

The presentation of the results follows the sequence of research questions listed above. Besides a *t*-test for group comparisons, the quality of the trained classifier resulting from data-mining and machine-learning procedures will be illustrated by using standard big data indicators, namely, Receiver Operating Characteristic (ROC) curves (Fawcett, 2006) and confusion matrices (Congalton, 1991).

RQ 1: Are there emotional reactions and indicators for cognitive load to be found that precede errors?

To answer this question, the previously described method was used to train a classifier on the whole training set features. The genetic programming reveals that the best result can be achieved with a gradient boosting classifier (Friedman, 2001). The corresponding ROC curve (see Figure 7) shows the performance of the trained classifier. The ROC curve visualises the diagnostic capability of a classifier: Therefore, the true positive rate (sensitivity; i.e. an error is correctly predicted) is plotted against the false positive rate (probability of false error predictions) while varying the different thresholds of the classifier. The finally received operating characteristic of the trained classifier can be seen in Figure 7.

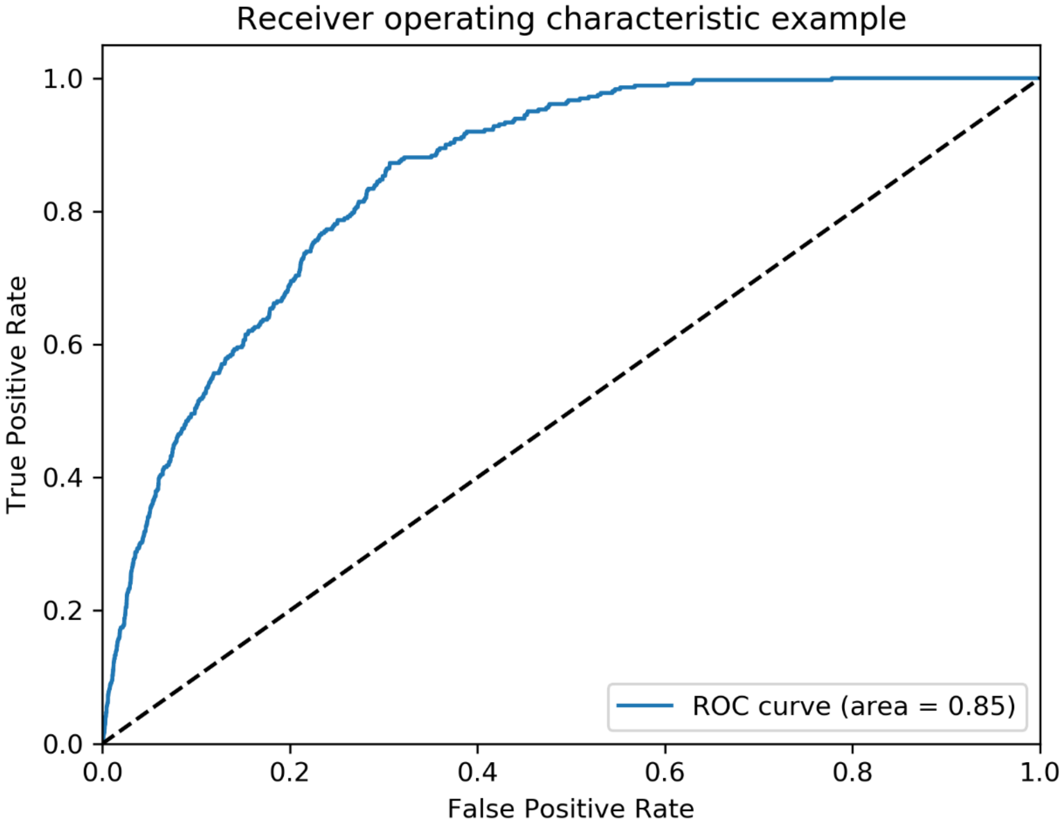


Figure 7. ROC curve of the optimal classifier.

The dashed line represents the baseline of random guessing (i.e. the pure chance of a wrong or correct error prediction is 0.5). An optimal classifier (i.e. each prediction is correct) would receive an ROC curve with an area of 1.0. The trained classifier here received a ROC curve with area of 0.85. Hence, with the current training set, it was possible to train a classifier able to precede errors sufficiently.

Changes in pupil diameters were considered as an indicator of cognitive load. The 300ms timeframes before errors and those before successes were examined and linear regressions for the changes in the test persons' pupil diameters before errors and before successes were calculated for each test person individually. Consequently, all test persons' beta-coefficients can be put into a *t*-test for independent samples distinguishing errors and successes. Table 2 presents the results of this *t*-test.

Table 2

Two-sided t-test comparing changes in pupil diameters before errors and successes.

	Mean	SD	*T*	*p*	*d*
Successes	0.000692	0.00301	2.57	.0103	0.135
Errors	0.00107	0.00258			

Note. df= 872.61.

In mean, changes in pupil diameter in error situations were significantly larger than changes in success situations.

RQ 2: Is the quality of learning related to emotional reactions?

Since the quality of learning was operationalised through the formula developed above, a median split was applied to distinguish two groups of test persons: better learners and worse learners. For this calculation, only those *n* = 19 test persons could be considered who finished level 6 in the game and whose face recognition was successful. On this basis, a *t*-test reveals differences in emotional reactions (i.e. *engagement* and *valence*).

Table 3

Two-sided t-test between better and worse learners.

	Better learners Mean (SD)	Worse learners Mean (SD)	*T*	*p*	*d*
Engagement*	.317 (.334)	.055 (.087)	-2.396	.028	1.100
Valence**	.232 (.355)	-.009 (.035)	-2.136	.048	1.303

Notes. * range [0,1]; ** range [-1,1].

Table 3 shows the results of the *t*-test that confirm the theoretical expectations: Better learners show significantly higher emotional reactions in terms of engagement and valence. This means that better learners show emotional reactions of a stronger amount or intensity than worse learners do, and they also tend to a higher extent towards positive emotions than worse learners do.

RQ 3: Are there appropriate online measures that indicate emotional reactions?

To answer this research question, the features that were used for training the error classifier can be analysed in more detail. An importance ranking of all features was calculated, based on chi-square statistical analyses between each feature and class. A chi-square analysis test is able to measure dependence between stochastic variables. For classification, this test can be used to obtain a measure of dependence between the features and the two classes of the classifier. The features that are most likely to be independent of class receive a low score and the features that are most likely to be dependent of class receive a high score. The more dependent a feature of the class is, the better this particular feature is for use in classification – in our case, for predicting an error. Figure 8 shows the 20 most important features for the prediction of errors.

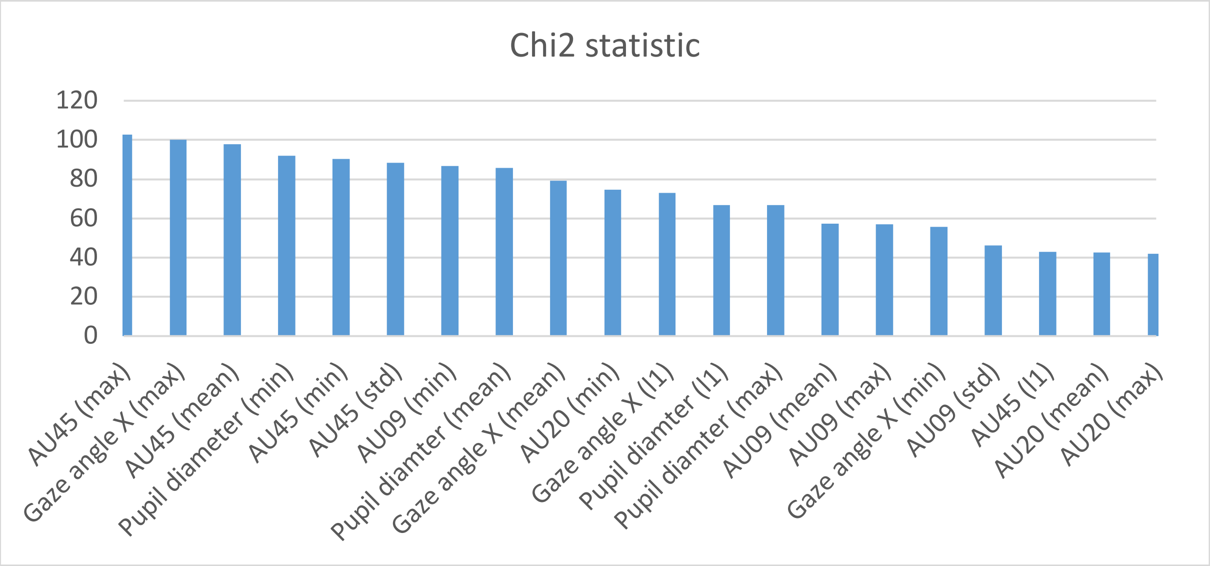


Figure 8. Ranking of most important features for the prediction of errors.

It is remarkable here that only eye blink (AU45), gaze (gaze angle x), cognitive load (pupil diameter), nose wrinkle (AU09) and lip stretcher (AU20) are represented in this top ranking. This means that those five items bear the majority of information required for the classification of errors.

This calculation of the importance of singular features for the prediction of errors reveals that not only can emotional reactions be considered as relevant precursors of errors but also that pupil diameters can be interpreted as indicators of cognitive load (i.e. significant changes within 300 milliseconds before an error occurs). Hence, a combination of pupil diameter features and facial video features contributes to the improvement of the classification of errors. In order to assess the increase in quality through this combination, the described machine pipeline was trained in two different ways: Option 1 considers all 168 features and Option 2 considers all features except the 6 pupil diameter features. Figure 9 shows confusion matrices for both options (Option 1, left side; Option 2, right side).

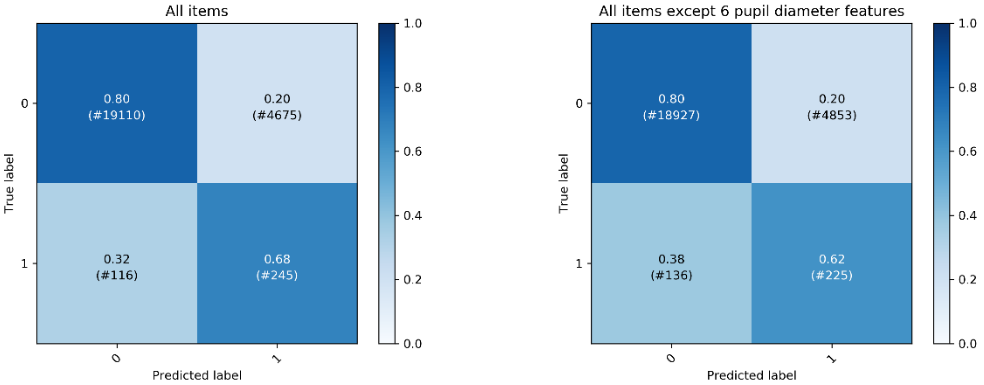


Figure 9. Confusion matrices.

These confusion matrices cover a four-field table. The x-axis marks the prediction of an event (0 = no error prediction; 1 = error prediction) and the y-axis marks the actual outcome of an event (0 = no error; 1 = error). Hence, the first and fourth quadrants indicate correct predictions, while the second and third indicate incorrect predictions. Hence: The upper left quadrant refers to true negative predictions, the upper right quadrant to false positive ones, the lower left quadrant to false negative ones, and the lower right quadrant to true positive predictions. The figure also comprises the absolute number of cases (#) and the probability of correct/incorrect predictions.

The comparison of both options reveals that the prediction of the first quadrant (i.e. the correct prediction of no error occurring) and the fourth quadrant (i.e. the correct prediction of an error) are slightly better if the pupil diameter features are also considered. In addition, the probability of detecting an error (prediction of no error occurring but error occurs – third quadrant) can be improved (from 38% down to 32%) if pupil diameters are used as well. Hence, considering pupil diameter features in addition to facial expression features slightly improves the overall score and the practical usage of the entire classification. The better performance can also be seen in the ROC analysis. The ROC curve of the classifier using all features can be seen above (figure 7). The ROC curve of the classifier excluding the pupil diameter features is shown in Figure 10.

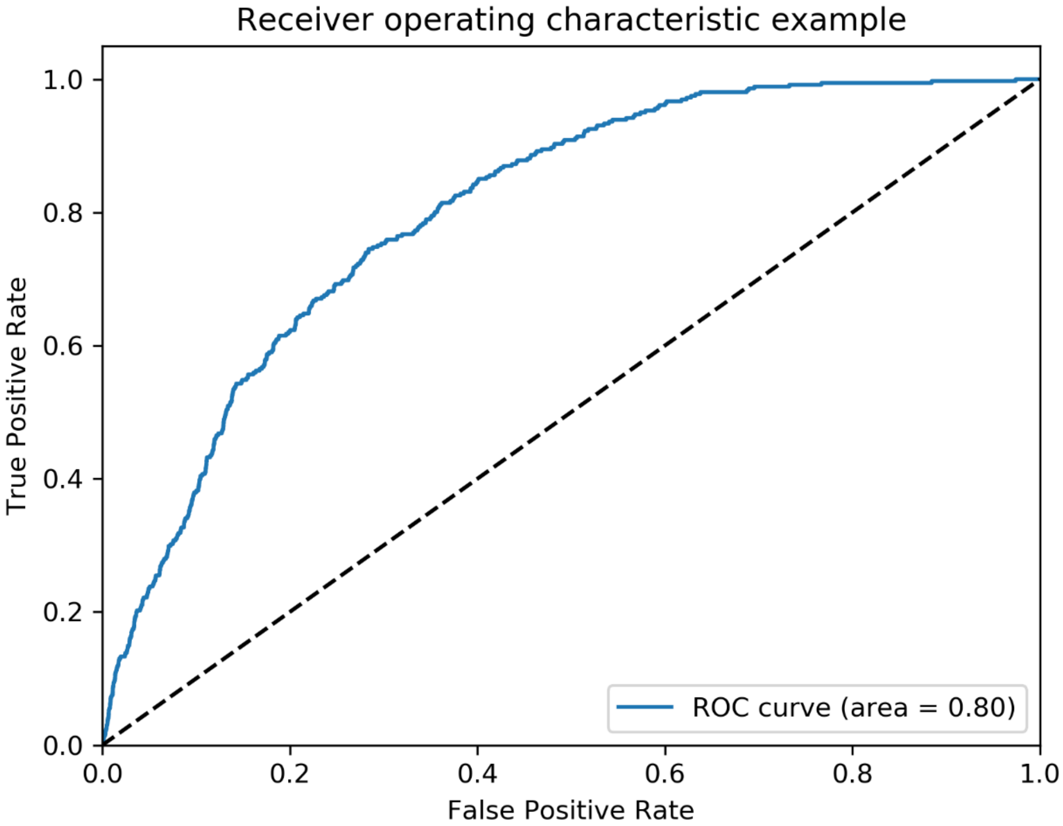


Figure 10. ROC curve for classifier excluding pupil diameter features.

The ROC curve area for the classifier that excludes pupil diameter features is 0.80, whereas the ROC curve area for the classifier considering all features is 0.85 (see Figure 7).

For the purpose of this study, this combination of indicators resulting from online measurement can be considered an appropriate measure of emotional reactions. It is important to emphasise that we did not strive to distinguish between different emotions but were simply interested in any kind of emotional reaction.

RQ 4: Is there a specific measure for the quality of learning from errors?

As the computer game provides similar tasks of increasing difficulty, it is plausible to assume that test persons improve by time, trials and errors. As a crucial measure of the change in performance (i.e. learning), the addition of angles α and β of the learning curve was chosen (see Figure 6). Table 3 reveals the descriptive statistics of the test persons' performance.

Table 4

Descriptive statistics for learning.

	N	Mean	SD	min	max
Learning	19	-1.75	14.33	-35.54	33.69

The test persons varied substantially in their performances. In total, four test persons scored positively, ten test persons scored negatively and five test persons scored zero.

For this particular setting, the learning curve indicates if and how an individual improves or develops during the course of a computer game of increasing difficulty. It indicates successful and unsuccessful attempts and thus can be considered a measure of the quality of learning from errors during the game.

5. Critical reflection on data quality and validity

The study comprises two variables: Learning from errors and emotional reactions. The discussion and critical reflection on data quality and validity will focus on each variable separately. Finally, a reflection follows on the relevance of data generated by a computer game for gaining insight into learning from errors.

5.1 Data on learning from errors

The construal of learning from errors follows quite a specific approach: The quality of learning – indicated through a learning score – was constructed as individual development along several trial-and-error attempts within a regular jump-and-run computer game.

The scores on learning appear to tend towards the negative side (see negative mean within Table 3), which requires a careful interpretation and must not be confused with a decrease in knowledge or negative learning. On the one hand, and as described above, the angles that result in the learning score were deliberately chosen because they refer to moments when the game's difficulty increased substantially (at levels 2 and 6). It is part of a regular performance to fail initially with an increase of difficulty and then to adapt to the new learning. Such regular performance results in a negative turn in the learning curve.

On the other hand, only those test persons could be considered for the analyses who were able finally to complete this level of increased difficulty. This implies that even the test person with the lowest learning score in the sample was able to master this task of increased difficulty – albeit while making the highest number of errors within the sample.

This measurement of learning faces two major limitations. First, test persons failing to master level 6 within the limited time could not be included in the measurement of the learning curve because the relation of successes and failures within level 6 could be determined only if a test person completed this level successfully within the given 5 minutes. Such a time-based cut-off results in a flawed learning curve for the last level. Future research attempts may allow as much time as a test person needs to cope with the increased difficulty. Second, very good test persons who master all challenges without any failures show a learning score of 0, because the learning score reflects individual development. Test persons who perform consistently well and thus do not show any difference in their error and success rates between the different levels do not develop in the sense of the measurement applied here. A learning score of 0 can be considered a ceiling effect because the task was not challenging enough for these test persons.

Hence, the data on learning from errors here indicate individual development during the run of a computer game which presumed that test persons fail occasionally during the run of the game. Individuals who performed constantly at the same (high or low) level received a learning score of 0. The learning score does not therefore provide information about the quality of performance but about the quality of individual development; that is the crucial aspect of learning from errors in the context of this setting and the important focus of the online data used here. Of course, the data would provide additional potential for focusing on learning from errors by directly connecting incidents when a test person initially fails and later succeeds during the game. However, since the focus of the study was on exploring opportunities to predict errors before they occur, the decision was made to include as many data as possible for the machine training. Considering combinations of initial failures and subsequent successes would have decreased the number of observed cases dramatically. In addition, there would also be alternative data available that reflect a test person's quality of performance (e.g. score, time, number of successes or failures), but such information does not tell us anything about learning from errors.

5.2 Data on emotional reactions

Without doubt, the face is an important means of expressing emotional reaction. The recorded video data made it possible to identify a variety of AUs that indicate emotional reactions. These indicators are based on analyses of fixpoints based on computer algorithms. In real face-to-face communications, the human mind is capable of processing the fine nuances of facial expressions unconsciously in order to interpret reactions appropriately. However, observational studies with human observers would probably not be able provide those kinds of data reliably. Hence, the kind of facial analyses provided in this study can be considered to be an advance.

This study – as already mentioned – did not aim at differentiating between different kinds of emotional reaction, however. This can be seen as a limitation, but for the context of the research questions raised here, this limitation does not have an impact on the data quality either for analysing learning from errors or for predicting errors because negative emotional reactions (e.g. fear) can limit human behaviour and situational perception in a similar way to positive reactions (e.g. euphoria).

As the results reveal, the features considered here for developing a classifier for errors are sufficient to predict an error before it occurs – in the context of the video game that was part of the investigation. The choice of features resulted from an exploratory procedure of data mining that searched for relevant patterns in the training set and then confirmed the choice in the test set of the sample. It is difficult to judge if the resulting values of a 70% correct error prediction, 80% correct prediction of non-occurring errors and an ROC curve of 0.85 are sufficient to justify applying these instruments in the technical context of man-machine-interaction. The acceptability of values probably depends on the range of application (e.g. high security areas or back-up systems). However, given that the classifier quite often predicted an error even though no error occurred (20%, in total > 4,500 cases, see Figure 8) it would still seem inappropriate for application in real contexts. One explanation might be that test persons showed emotional reactions but still managed to avoid making an error in the video game. Further machine-learning procedures may help to reduce this kind of wrong prediction.

5.3 Relevance of data generated through a computer game

Learning from errors during a computer game may be seen as very different from learning from errors in real life, particular workplace settings. Indeed, learning from errors at work occurs within an organisational error culture (Putz, Schilling, & Kluge, 2012) that cannot be transferred to a laboratory setting. However, learning during work and learning during a computer game share important similarities: In both situations, learning occurs as a by-product of the intention to reach a goal. In both cases, there is no curriculum and no instruction that guides the acting but simply the intention to reach the goal successfully. Learning from errors in real-life contexts varies widely across occasions and individuals. Hence, it is difficult to identify the general characteristics of learning from errors empirically because situations are not comparable enough. A computer game, by contrast, provides stable conditions across all test persons and makes it possible to observe learning processes by looking at the tasks at hand. It should therefore provide enough insights into general processes related to learning from errors that can claim relevance to learning from errors in real-life contexts.

6. Conclusions

First, the findings reveal on the one hand that non-specified emotional reactions antecede test persons' failures to overcome an obstacle or avoid an attack. Second, the findings reveal that test persons who show a beneficial pattern of emotional reactions – that is, a higher extent of engagement and a positive valence – achieve higher learning scores than do test persons with an awkward pattern of emotional reactions. Hence, they confirm Oser's theory about the importance of emotions for learning from errors (Oser & Spychiger, 2005; Oser & Volery, 2012). For the field of learning from errors in working life, this finding reveals the importance of the organisational culture (Schein, 2004) and team climate in workplaces (Edmondson, 1999). These require the appropriate social conditions that accept emotional reactions without generating disadvantages for the failing person. Conditions that fail to provide such an environment tend to provoke the concealment, disregard and thus repetition of errors (Marsick & Watkins, 2003).

On the other hand, given the extent to which the results fit into theoretical patterns, the findings also indicate that the tested way of gathering online data is a promising one. Certainly, the procedures applied in this study have potential for improvement, as discussed above. As long as there are no repeat studies applying the same or similar measures, we do not know much about the validity of such measures. The experiences from this study suggest the need to repeat it under improved circumstances in two respects. First, a time limit is to be avoided; all test persons should receive as much time as they require to master level 6 of this game – as long as it appears reasonable to expect that each test person is able to cope with the difficulties of level 6 within a reasonable time. Second, a repeat of this study should use a larger sample that would make it possible to connect initial failure with subsequent success directly in order to permit a focus on and analysis of concrete cases of learning from errors. In addition, a repeat of this study with a larger sample would provide an appropriate set of data to test the quality of the classifier found herein.

Keypoints

- Emotional reactions and cognitive load precede errors.
- Measures of automated face recognition generate data coherent with literature on emotions.
- The combination of face recognition and eye-tracking data can be used to predict errors before they occur.
- Online measurements confirm keypoints of the theory of learning from errors.

References

Baltrusaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace. An open source facial behavior analysis toolkit. In WACV (Ed.), *2016 IEEE Winter Conference on Applications of Computer Vision* (pp. 1-10). Lake Placid: IEEE. doi: 10.1109/WACV.2016.7477553

Bauer, J., & Mulder, R. H. (2013). Engagement in learning after errors at work: Enabling conditions and types of engagement. *Journal of Education and Work, 26*(1), 99–119.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276-292.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35-46.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, *38*(2), 105-134.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350-383.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A technique for the measurement of facial movements*. Sunnyvale: Consulting Psychologist Press.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232.

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning: Studies in Vocational and Professional Education, 1*(2), 87–103.

Harteis, C., & Bauer, J. (2014). Learning from errors at work. In S. Billett, C. Harteis & H. Gruber (Eds.), *International handbook of research in professional and practice-based learning* (pp. 699-732). Dordrecht: Springer Academics.

Harteis, C., Bauer, J., & Gruber, H. (2008). The culture of learning from mistakes: How employees handle mistakes in everyday work. *International Journal of Educational Research, 47*(4), 223–231.

Hetzner, S., Gartmeier, M., Heid, H., & Gruber, H. (2011). Error orientation and reflection at work. *Vocations and Learning: Studies in Vocational and Professional Education*, *4*(1), 25-39.

Hofman, D. A., & Frese, M. (2011). Errors, error taxonomies, error prevention, and error management: Laying the groundwork for discussing errors in organisation. In D. A. Hofmann & M. Frese (Eds.), *Errors in organisations* (pp. 1–43). London: Routledge.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking. A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Jaber, M. Y. (Ed.). (2016). *Learning curves: Theory, models, and applications*. Boca Raton: CRC Press.

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE*, *13*(9), e0203629.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry. A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18-27.

Marsick, V. J., & Watkins, K. E. (2003). Demonstrating the value of an organization's learning culture: The dimension of the learning organization questionnaire. *Advances in Developing Human Resources*, *5*(2), 132-151.

McDuff, D., El Kaliouby, R., & Picard, R. W. (2015, September). Crowdsourcing facial responses to online videos. In IEEE (Ed.), *Affective Computing and Intelligent Interaction (ACII), 2015* (pp. 512-518). Piscataway township: IEEE.

Olson R.S., Urbanowicz R.J., Andrews P.C., Lavender N.A., Kidd L.C., & Moore J.H. (2016). Automating biomedical data science through tree-based pipeline optimization. In G. Squillero & P. Burelli (Eds.), *Applications of Evolutionary Computation. EvoApplications 2016. Lecture Notes in Computer Science* (pp. 123-137). Cham: Springer.

Oser, F., Müller, S., Obex, T., Volery, T., & Shavelson, R. J. (2018). Rescue an enterprise from failure: An innovative assessment tool for simulated performance. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A., C. Lautenbach & C. Kuhn (Eds.),*Assessment of learning outcomes in higher education* (pp. 123-144). Springer, Cham.

Oser, F., Näpflin, C., Hofer, C., & Aerni, P. (2012). Towards a theory of Negative Knowledge (NK): Almost-mistakes as drivers of episodic memory amplification. In J. Bauer & C. Harteis (Eds.), *Human fallibility. The ambiguity of errors for work and learning* (pp. 53–70). Dordrecht: Springer.

Oser, F., & Obex, T. (2015). Gains and losses of control: the construct "Sense of Failure" and the competence to "Rescue an Enterprise from Failure". *Empirical Research in Vocational Education and Training*, *7*(1), 3.

Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft*. Beltz: Weinheim.

Oser, F., & Volery, T. (2012). *"Sense of Failure" and "Sense of Success" among entrepreneurs: the identification and promotion of neglected twin entrepreneurial competencies*. Bern: SKBF.

Putz, D.,Schilling, J., & Kluge, A. (2012). Measuring organizational climate for learing from errors at work. In J. Bauer & C. Harteis (Eds.), *Human fallibility* (pp. 107-123). Dordrecht: Springer.

Rybowiak, V., Garst, H., Frese, M., & Batinic, B. (1999). Error orientation questionnaire (EOQ): Reliability, validity, and different language equivalence. *Journal of Organizational Behavior*, *20*, 527-547.

Schein, E. H. (2004). *Organizational culture and leadership*. San Francisco: Jossey-Bass.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295-312.

Szulewski, A., Kelton, D., & Howes, D. (2017). Pupillometry as tool to study expertise in medicine. *Frontline Learning Research*, *5*(3), 55-65.

Wolf, K. (2015). Measuring facial expression of emotion. *Dialogues in Clinical Neuroscience*, *17*(4), 457-462.

Yelle, L. E. (1979). The learning curve: Historical review and comprehensive survey. *Decision Sciences*, *10*(2), 302-328.

Zaichkowsky, J. L. (1994). The Personal involvement Inventory: Reduction, revision, and application to advertising. *Journal of Advertising*, *23*(4), 59-70.