Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Stolicyn A, Steele JD, Seriès P (2022). Prediction of depression symptoms in individual subjects with face and eye movement tracking. *Psychological Medicine* **52**, 1784–1792. https://doi.org/10.1017/S0033291720003608

Received: 3 January 2020
Revised: 13 August 2020
Accepted: 11 September 2020
First published online: 9 November 2020

Key words:
Cognitive tasks; depression; eye movements; eye-tracking; face movements; machine learning; prediction

Author for correspondence:
Aleks Stolicyn
E-mail: a.stolicyn@ed.ac.uk

Prediction of depression symptoms in individual subjects with face and eye movement tracking

Aleks Stolicyn[1,2], J. Douglas Steele[3] and Peggy Seriès[2]

[1]Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Kennedy Tower, Royal Edinburgh Hospital, Morningside Park, Edinburgh EH10 5HF, UK; [2]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK and [3]Division of Imaging Science and Technology, School of Medicine, Dundee University, Ninewells Hospital & Medical School, Dundee DD1 9SY, UK

Abstract

Background. Depression is a challenge to diagnose reliably and the current gold standard for trials of DSM-5 has been in agreement between two or more medical specialists. Research studies aiming to objectively predict depression have typically used brain scanning. Less expensive methods from cognitive neuroscience may allow quicker and more reliable diagnoses, and contribute to reducing the costs of managing the condition. In the current study we aimed to develop a novel inexpensive system for detecting elevated symptoms of depression based on tracking face and eye movements during the performance of cognitive tasks.
Methods. In total, 75 participants performed two novel cognitive tasks with verbal affective distraction elements while their face and eye movements were recorded using inexpensive cameras. Data from 48 participants (mean age 25.5 years, standard deviation of 6.1 years, 25 with elevated symptoms of depression) passed quality control and were included in a case-control classification analysis with machine learning.
Results. Classification accuracy using cross-validation (within-study replication) reached 79% (sensitivity 76%, specificity 82%), when face and eye movement measures were combined. Symptomatic participants were characterised by less intense mouth and eyelid movements during different stages of the two tasks, and by differences in frequencies and durations of fixations on affectively salient distraction words.
Conclusions. Elevated symptoms of depression can be detected with face and eye movement tracking during the cognitive performance, with a close to clinically-relevant accuracy (~80%). Future studies should validate these results in larger samples and in clinical populations.

Introduction

At present, depression is diagnosed by medical practitioners in both primary and secondary medical care settings. The diagnostic criteria are subjective: the patient's symptoms are evaluated during a clinical interview (American Psychiatric Association, 2013; National Institute for Health and Care Excellence, 2009). Major depressive disorder (MDD) remains a challenge for reliable diagnosis and evidence indicates a relatively low rate of diagnostic agreement between specialists (Freedman et al., 2013). This is in part because MDD ranges from mild illness which merges with normal experience, to moderate illness, to gravely ill patients, while the DSM (Diagnostic and Statistical Manual of Mental Disorders) does not capture this distinction. MDD also often co-occurs with anxiety and this can cause diagnostic confusion. Over the past decade, many studies have attempted automated diagnostic classification of depression in standardised settings with *machine learning* methods and neuroimaging data. Very good results – depression detection accuracies up to 90% – have been achieved with brain structural measures (Johnston, Steele, Tolomeo, Christmas, & Matthews, 2015; Mwangi, Ebmeier, Matthews, & Steele, 2012), brain functional connectivity measures (Wei et al., 2013; Zeng et al., 2012), and task-related activation measures (Johnston, Tolomeo, et al., 2015; Rosa et al., 2015). Despite generally promising results (review in Kambeitz et al., 2017), brain scanning remains logistically expensive and requires technical expertise, which could limit translation of findings to clinical practice (online supplementary section S1.3). In contrast to neuroimaging, some behavioural aspects of depression ('signs') can be relatively inexpensive to measure. Studies which applied *facial electromyography* (EMG) (e.g. Gehricke & Shapiro, 2000; Rottenberg, Gross, & Gotlib, 2005) or *manual face movement ratings* (e.g. Renneberg, Heyn, Gebhard, & Bachmann, 2005; Sloan, Strauss, Quirk, & Sajatovic, 1997) indicate altered eyebrow, cheek and mouth movements when imagining or viewing affective materials (pictures, scenery or clips) – although the direction of change could be dependent on the experimental conditions and the participant sample.

CAMBRIDGE
UNIVERSITY PRESS

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press

Automated video-based facial behaviour analysis methods have also been applied in several studies to characterise behaviour in clinical interviews (e.g. Girard *et al.* 2014; Stratou, Scherer, Gratch, & Morency, 2015), but these methods have not yet been used to study reactions to affective material. Changes in eye movements have also been reported, with evidence indicating that patients with depression fixate more often and for longer on negative affective materials and less on positive materials (Armstrong & Olatunji, 2012; Carvalho et al., 2015). Several studies have also been successful at the diagnostic classification of depression with eye movement measures (e.g. Alghowinem, Goecke, Wagner, Parker, & Breakspear, 2013) or with automated face movement analysis (review in Pampouchidou et al., 2019), but only in the context of clinical interviews. In the present study, we aimed to develop an *inexpensive* system for detecting signs of depression based on combined face and eye movement tracking during cognitive performance with affective distractions. We designed two novel cognitive tasks with affective distraction elements and recorded face and eye movements from a cohort of young non-clinical participants, with or without elevated symptoms of depression – when they performed the cognitive tasks. We then applied a machine learning technique to assess how well the recorded face and eye movement measures could discriminate between symptomatic and non-symptomatic participants. We hypothesised that classification with combined face and eye movement measures may perform better compared to either modality separately. We provide additional rationale for investigating face and eye movement for the detection of depression in section S1 of the online supplementary material.

Methods

Experiment participants

A total of 75 participants were recruited mainly from the student and recent graduate population at the University of Edinburgh. General participant requirements included having normal or corrected to normal vision and either being a native English speaker, or having lived in a mainly English-speaking country and using English as the primary language for the past 7 years. Recruitment criteria for *symptomatic* participants included low mood and/or loss of interest in daily activities over the past two weeks. Each participant was paid £15 for their participation, which took up to 1.5 hours.

In the first part of the experiment, participants reported their age and caffeine consumption, and then completed Alcohol Use Disorders Identification Test (AUDIT) and Center for Epidemiologic Studies Depression Scale (CES-D) questionnaires, as well as the National Adult Reading Test (NART) (Bright, Jaldow, & Kopelman, 2002; Radloff, 1977; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993). Caffeine consumption was measured in cups of coffee per day, where one cup of tea was assumed to be equal to half a cup of coffee. Caffeine was measured primarily because it may have an effect on cognitive performance (McLellan, Caldwell, & Lieberman, 2016), although some evidence also indicates that it may be associated with a lower risk for depression (Wang, Shen, Wu, & Zhang, 2016). Participants were classed as symptomatic if they scored strictly above the threshold of 16 in CES-D, and non-symptomatic (control) if they scored strictly below the threshold. Three participants who had the CES-D score exactly at the threshold were excluded from the analyses. CES-D was applied because it is a freely available tool designed to screen for depression symptoms in general population samples such as in our study.

All participants were informed about the course and content of the experiment and provided informed consent. All procedures in the study complied with the Declaration of Helsinki of the World Medical Association, as revised in 2013. The experiment was approved by the School of Informatics Ethics committee at the University of Edinburgh.

Technical setup

The technical setup consisted of a desktop computer, a screen with a keyboard, an eye-tracking device, and a digital camera. A black wall-screen was positioned approximately 1.5 metres behind the participant to provide background for visual recordings.

Screen and response capture

Experimental stimuli were presented on a 21.5 inch (54.6 cm) screen. Participants sat approximately 60 cm to 90 cm from the screen as they felt comfortable. Responses during cognitive tasks were captured using a Hewlett Packard keyboard with four distinctively marked response keys. It should be noted that most consumer keyboards cannot provide millisecond-level timing accuracy and typically have variable response delays between 15 and 40 milliseconds (Plant & Turner, 2009). Response times captured during the experiment were hence accurate only to a limited degree.

Face movement recordings

Digital visual recordings were made with an Intel RealSense SR300 camera (Intel Inc.). The camera was positioned directly on top of the screen and captured the participants' faces. Recordings were made at a resolution of 1280 × 720 with a stable frame rate of 30 frames per second. One block-synchronised recording was made per block of trials for each task. Prior to starting the first cognitive task, participants were asked to keep a neutral face expression for approximately 10 s for a recording of baseline facial expression. This was then used for correction during the analysis stage.

Eye movement tracking

Eye-tracking data were collected using a Gazepoint GP3 eye-tracker (Gazepoint, Canada; Zugal and Pinggera, 2014), which was positioned directly below the screen. Eye-tracking data were sampled at a 60 Hz rate, with the advertised device accuracy between 0.5° and 1°, spatial resolution of 0.1°, and up to 50 ms tracking latency. Gazepoint GP3 remains one of the most accessible and least expensive eye-tracking devices available on the market as of 2019. The device was adjusted manually to each participant's height and five-point calibration was performed before the first cognitive task and between trial blocks if it was necessary due to the participant's head movements. Eye-tracking measures for the different visual elements of the tasks were captured in real-time during task performance, with a 40 ms correction to account for tracking latency.

Cognitive tasks

Two cognitive tasks with affective distractions were developed as part of the study. The Delayed Match to Sample (DMS) task probes working memory, whilst the Rapid Detection (RD) task assesses sustained attention. These two cognitive domains have been reported as compromised in depression (McIntyre et al., 2013; Rock, Roiser, Riedel, & Blackwell, 2014). The number of trials in the tasks was defined to limit the total assessment time to approximately 50 minutes or less.

Delayed match to sample task

During the DMS task, participants were required to memorise an initial sample pattern and then identify it among four alternatives, after a brief delay. Each pattern in the task consisted of four coloured quadrants with variable numbers of white marks in each quadrant. During the 12-second delay stage at each trial, four words flashed at different locations on the screen to distract the participant from their task. An additional distraction word was displayed at the response stage alongside the four patterns. Feedback was given to the participant at each trial to indicate a correct or incorrect response. Task trial structure is illustrated in Fig. 1*a* and further details can be found in online supplementary section S2.1.

There were three blocks of trials in the task, with 15 trials in each block (45 trials in total). The first block of trials had neutral distraction words, the second had positive words, and the third featured negative words. The block with negative words was always presented last to avoid carry-over of any effects of negative words between the blocks. Each participant was allowed to have brief breaks to rest between trial blocks. A training sequence with four neutral trials was administered before the first block of trials. Task instructions were read by the experimenter and also appeared on screen in the text before the training block.

Rapid detection task

At each trial of the RD task, participants were first presented with a target letter (target stage), and then required to identify it, by pressing the space key, among 44 letters which flashed on the screen one after another (detection stage). Five target letters were distributed uniformly among others, with the first four letters always non-target. Each letter flashed initially, and then faded away for 1000 ms. Participants received feedback when they correctly detected the target letter, when they missed a target letter, or when they made an erroneous response to a non-target letter. Throughout the detection stage, five words appeared alongside the flashing letters, one at a time, to distract the participant from their task. Five distraction words also flashed at the centre of the screen between the target and detection stages, again to distract participants from the task. RD task trial structure is illustrated in Fig. 1*b* and further details can be found in online supplementary section S2.2.

As with the DMS task, there were three blocks of trials – neutral first, positive second, and negative third. Each block consisted of five trials, with breaks for rest between the blocks (15 trials in total). A single training trial was administered prior to the first block, together with task instructions. Instructions were read by the experimenter and were also displayed on the screen. Participants were explicitly asked to try and look at each distraction word at least once. If a participant did not look at least at one of the distraction words (according to the eye-tracking readings), a feedback message was displayed at the end of the trial, reminding the participant to try and look at each word. In cases where eye-tracking was inaccurate according to calibration results, participants were notified that feedback may be incorrect.

Affective distractions

In total, 60 neutral, 60 positive and 60 negative distraction words were selected from the Warriner database of 13 915 English lemmas (Warriner, Kuperman, & Brysbaert, 2013). Section S2.3 in the online supplementary material outlines further details on the selected distraction words.

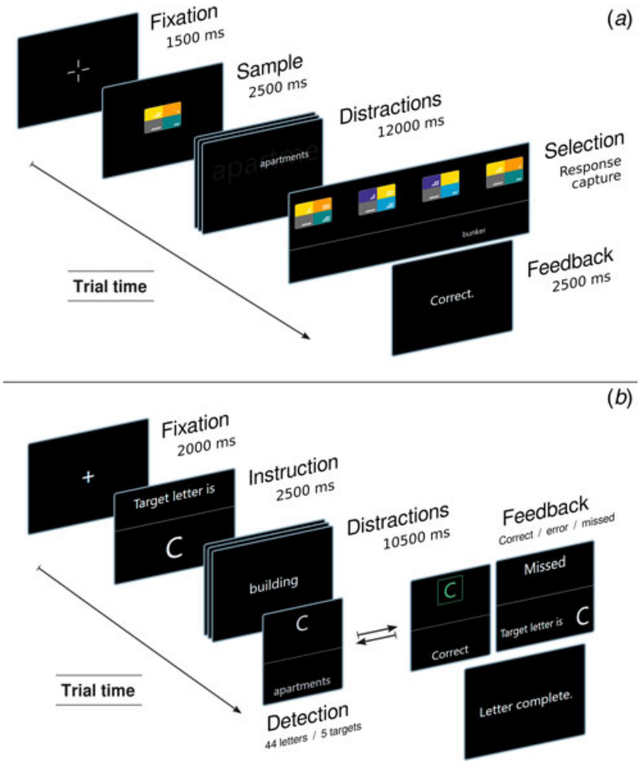

Fig. 1. Single-trial timelines for the DMS task (*a*) and for the RD task (*b*).

Recorded measures

Behavioural measures

Mean reaction times (RTs) for correct and error responses, as well as *accuracies* (correct response rates) were computed for the entire DMS task (45 trials). *Mean RTs* for correct and error responses were also computed for the entire RD task (15 trials). *Detection rates* were calculated as percentages of correctly detected target letters. *Error counts* were computed as numbers of erroneous responses.

Face movement measures

Recordings were segmented (epoched) into time-locked parts related to seven trial stages of the DMS task and six trial stages of the RD task (i.e. sample, distraction, selection, feedback and others). For each epoch, time series of intensities (one measurement per frame, scale from 0 to 5) were extracted for 17 facial action units (AUs) from the Facial Action Coding System (FACS, Ekman, Friesen, & Hager, 2002), using OpenFace toolkit (Baltrusaitis, Mahmoud, & Robinson, 2015, 2016). AU intensity time series were baseline-corrected using mean AU intensities from the participant's 10-second baseline recording. After baseline correction three metrics of interest were extracted for each AU in each epoch: (1) *maximal AU intensity*, (2) *average AU intensity above threshold*, and (3) *duration of AU above threshold*. The threshold value was here set to 1 (on the scale from 0 to 5). The second and third metrics correspond to average intensity and duration of AU activity in the epoch. Means of the

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press

metrics for each AU in each of the 13 task stages were then calculated, which resulted in 663 facial behaviour measures for each participant (3 metrics × 17 AUs × 13 task stages = 663 measures). AU durations were represented in seconds and AU intensity measures were on the scale from 0 to 5. Missing face-tracking measures – for example when the participant did not make any errors at the DMS or RD tasks – were replaced with value −1. A diagram illustrating facial movement measure extraction can be found in Figure S3 in the online supplementary material.

Eye movement measures

Three eye-tracking metrics were recorded for 17 visual elements in each DMS trial and 15 visual elements in each RD trial. These metrics were: (1) *latency of the first eye fixation* (since the appearance of the element), (2) *count of fixations*, and (3) *total time when fixated on the element*. For each participant, the metrics were averaged across trials, which resulted in 51 measures for the DMS task and 45 measures for the RD task. In addition, differences between metric means in positive and neutral conditions, as well as negative and neutral conditions were calculated, resulting in 24 further measures for the DMS task and 12 further measures for the RD task. Tables S1–S4 in the online supplementary material list eye-tracked elements and metric difference measures for the two tasks.

Classification methods

Feature selection

Overall, there were 663 facial movement and 132 eye-tracking measures for each participant. To improve classification results we performed feature selection using a simple statistical filter – two-sample *t* test with assumed unequal variances between samples (Welch's *t* test). Only features which were significantly different between the two classes in the training data at a specified *p* value threshold were selected for classifier training and testing at each cross-validation iteration (fold) (e.g. see Mwangi, Tian, and Soares, 2014). The *p* value threshold was optimised using grid search within a *nested* cross-validation scheme (section S4.1 in online supplementary material).

Classification model

Support vector machine (SVM) with a Gaussian (radial basis function) kernel was used as the classification model in the study (Cortes & Vapnik, 1995). SVM is the most frequently used classification technique in neuroimaging classification studies of depression (Kambeitz et al., 2017). The classifier has two hyperparameters – *regularisation* (box constraint) and *kernel scale*. The regularisation parameter was set to 1 and kernel scale parameter was set to 9 (section S4.2 in online supplementary material). Before classifier training and testing, features were standardised – centred and scaled by feature means and standard deviations in the training data. Classifier training and testing was performed with MATLAB R2018a Statistics and Machine Learning Toolbox (Mathworks Inc.). Alternative classification models were investigated *post hoc* and results of these analyses are described in sections S4.3 and S6 in the online supplementary material.

Cross-validation

Leave-one-out cross-validation (LOOCV) was used to assess the performance of the classification model for detecting elevated symptoms of depression. Briefly, at each iteration of LOOCV, one data sample (participant) is first excluded from the complete dataset. The classification model is then trained on the remaining data and tested on the excluded sample. This is repeated for each sample and test outcomes are averaged to define an overall predictive accuracy. We opted to perform LOOCV because it has been most widely used in the previous studies (Gao, Calhoun, & Sui, 2018), maximizes the amount of training data, and provides the least biased accuracy estimates (Zhang & Yang, 2015). Cross-validation can in general be interpreted as within-study replication, as the classification model is trained and tested repeatedly for each participant in the study.

Results

Behavioural performance

Participant sample

The behavioural task performance was assessed in 72 participants (34 symptomatic). Symptomatic and control groups were balanced with respect to gender (18 symptomatic and 18 control female participants). The difference in age approached significance ($p = 0.0564$, mean control age 25.7, mean symptomatic group age 23.4), but the two groups were not significantly different in NART or AUDIT scores. Symptomatic participants on average reported consuming ½ more cups of coffee per day ($p = 0.0225$). Summary demographic characteristics of the sample can be found in Table S5 in the online supplementary material.

Behavioural results

Mean accuracy at the DMS task for all participants was 90.0% (standard deviation 8.24%). Two-sample *t* tests did not reveal any significant differences in reaction times (correct or error) and accuracies between symptomatic and control participants. At the RD task participants on average detected 97.3% of target letters and made 1.6 errors. Two-sample *t* tests did not reveal any significant differences in reaction times between the groups. Differences in *detection rates* and *error counts* approached significance – symptomatic participants tended to detect on average 1.15% more target letters ($p = 0.057$), and tended to make on average 0.66 fewer errors ($p = 0.077$). The effects of depression symptoms on performance remained non-significant when controlling for *age* and *caffeine* consumption within additional one-way ANCOVA tests. Summary performance measures for the sample can be found in Tables S6 and S7 in the online supplementary material.

Case-control classification

Participant sample

Of 72 participants in total, 12 were excluded from classification analyses due to face-tracking problems (OpenFace analysis), 11 due to eye-tracking faults, and one participant due to problems with both. Briefly, eight participants were excluded due to problems in correctly tracking the chin or lower part of the face. Either upper or lower lips were not correctly localised for another four participants. Finally, for one participant there were problems in correctly tracking the left side of the face. With regard to eye-tracking faults, in six cases the eye-tracker could not stably or correctly localise either one or both participant's pupils. For one participant, calibration was inaccurate after several attempts. Two participants moved in and out of the eye-tracking camera

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press

Table 1. Characteristics of the sample used for classification analyses

	Group		
	Control	Symptomatic	*p* value
Size (male / female)	23 (12 / 11)	25 (12 / 13)	–
Age	27.5 (7.7)	23.7 (3.2)	$p = 0.035$
NART	35.2 (3.2)	36.9 (3.6)	N.S.
AUDIT	6.8 (5.1)	7.1 (7.0)	N.S.
Caffeine	1.0 (0.9)	1.3 (1.2)	N.S.
CES-D	8.4 (4.3)	25.6 (6.6)	$p < 0.00001$

Note: Caffeine is in cups of coffee per day. Standard deviations are in parentheses. *p* value defined according to two-sample independent *t* tests.

field-of-view during the assessment. Finally, the eye-tracker intermittently lost track of eyes due to reflections on participant glasses in another three cases. This resulted in a final sample of 48 participants included in the classification analysis.

Of the analysed 48 participants, 25 were symptomatic and 23 were controls. This sample size is consistent with most of the previous depression classification studies (Kambeitz et al., 2017). Symptomatic participants were on average 3.8 years younger than controls ($p = 0.035$), but there was no significant difference in other measures. Table 1 outlines characteristics of the sample used for classification analyses.

Classification results

SVM classification accuracy with combined *face-tracking* and *eye-tracking* features reached 79.17% (sensitivity 76%, specificity 82.61%). We attempted classification with features from each domain separately to check if combining both domains achieves the best results. Classification with only face-tracking features reached 66.67% accuracy (sensitivity 68%, specificity 65.22%). Classification with only eye-tracking features reached 64.58% accuracy (sensitivity 68%, specificity 60.87%). This indicates that face and eye movement measures complement each other to achieve the best results.

Classification features

Table 2 outlines the set of *face-tracking* and *eye-tracking* features selected in at least 80% of LOOCV folds (consensus features), with effect sizes and significance values calculated for the entire analysed sample (48 participants, Table 1). Of the 50 identified *face-tracking* consensus features, four were related to the DMS task and 46 to the RD task. Of the 11 eye-tracking features, seven were related to the DMS task and four were related to the RD task. All selected *face-tracking* metrics were reduced in symptomatic participants with medium effect sizes according to Cohen's *D* criteria, with an exception for mean intensity in AU9 (nose wrinkler), which was increased during the negative distraction stage of the DMS task. Since some participants completed the RD task without missing any target letters, effect sizes are not displayed for the metrics related to the RD missed-target feedback.

Of the 48 analysed participants, nine detected every target letter at the RD task. Face-tracking features for the RD *missed-target feedback* stage in these cases were replaced with value −1 (methods section, all features in the analysis are positive and most have numerical values between 0 and 5). The replaced features implicitly incorporate information about cognitive performance – i.e. whether the participant had detected every target letter or not. For classification purposes, this information replaced the missing facial movement data. A post-hoc χ^2 test confirmed that the symptomatic sample had a larger proportion of participants who detected every letter at the task (8 out of 25 symptomatic compared to 1 out of 23 controls, $\chi^2 = 6.013$, $p = 0.0142$). Classification based only on the 45 features related to the RD *missed-target feedback* achieved a 56.25% accuracy – this indicates that information about cognitive performance complemented information about face and eye movements to achieve a higher accuracy.

Discussion

Depressive symptom detection

Depression classification

To the best of our knowledge, our study is the first to assess the application of face and eye movement tracking during cognitive task performance for detection of elevated symptoms of depression. The results suggest that face and eye movement measures may be promising for future research and that the best accuracy can be achieved when these measures are combined.

Our study spans four out of five domains outlined in the Research Domain Criteria (RDoC). RDoC is a leading mental health research initiative supported by the US National Institute for Mental Health (Cuthbert, 2014). The initiative aims to define mental health conditions in terms of their characteristics grounded in biology and neuroscience, as compared to symptom-based definitions in the current diagnostic manuals (ICD-10 and DSM-5). RDoC proposes five domains relevant for mental health – *negative valence systems*, *positive valence systems*, *cognitive systems*, *social process systems*, and *arousal systems*. It is hoped that different mental health conditions and their subtypes can be defined by characteristics in these domains, leading to more objective diagnoses. Within our study, the tasks assessed the cognitive systems, whilst affective distractions aimed to probe the positive and negative valence systems. Moreover, facial movements are related to systems for social processes. Our results – classification accuracy close to 80% – support the assertion that diagnosis of depression could in principle be performed using behavioural measures related to these four RDoC domains.

Classification accuracy in our study was similar to those in the previous investigations with brain imaging data (Kambeitz et al., 2017), although lower compared to the most promising results for more severe depression (up to and above 90%, Johnston, Tolomeo, *et al.* 2015; Mwangi *et al.* 2012; Zeng *et al.* 2012). Further work should focus on the improvement of the technical setup and on the assessment of clinical participants – we briefly discuss these aspects below.

Technical design

One distinct advantage of our methods is the simplicity of the technical setup. In brain MRI studies, for example, participants have to undergo a scanning process, which is expensive and requires assistance from highly-trained radiographers. A T1-weighted brain scan, however, only takes around 7 minutes and NHS radiology departments have the necessary equipment (Steele & Paulus, 2019). In contrast, the technical setup in our study involved only a relatively inexpensive eye-tracking device

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press

Table 2. Identified cross-validation *consensus* face-tracking and eye-tracking features

Task	Task stage	Facial AUs	AU metric	Depressive symptom effect	Significance
DMS	Negative distraction	AU9	Mean intensity	0.670	$p = 0.0263$
	Match selection	AU17	Mean intensity	−0.648	$p = 0.0379$
		AU23	Max intensity	−0.745	$p = 0.0153$
	Correct feedback	AU5	Max intensity	−0.678	$p = 0.0236$
RD	Target instruction	AU15	Max intensity	−0.688	$p = 0.0215$
		AU17	Max intensity	−0.657	$p = 0.0287$
	Intro negative distraction	AU14	Max intensity	−0.783	$p = 0.0110$
	Correct feedback	AU14	Max intensity	−0.709	$p = 0.0201$
		AU17	Max intensity	−0.712	$p = 0.0174$
		AU26	Max intensity	−0.730	$p = 0.0163$
	Missed feedback	AU5 AU45	Active time	–	–
		AU26	Max intensity		
		AU4	Max intensity Mean intensity		
		AU2 AU15 AU20 AU25	Max intensity Active time		
		AU1 AU6 AU7 AU9 AU10 AU12 AU14 AU17 AU23	Max intensity Mean intensity Active time		
Task	**Task stage**	**Eye-tracked element**	**Eye-tracking metric**	**Depressive symptom effect**	**Significance**
DMS	Distraction	Positive word	Count of fixations	−0.706	$p = 0.0196$
		Positive to neutral word difference	Count of fixations	−0.780	$p = 0.0112$
	Match selection	Any distraction word	Fixation time	0.671	$p = 0.0257$
			Count of fixations	0.706	$p = 0.0193$
		Positive distraction word	Fixation time	0.663	$p = 0.0273$
		Negative distraction word	Fixation time	0.629	$p = 0.0355$
		Positive to neutral word difference	Fixation time	0.626	$p = 0.0358$
RD	Intro distraction	Negative to neutral word difference	Fixation time	−0.773	$p = 0.0109$
	Target detection	Neutral distraction word	Fixation latency	−0.682	$p = 0.0252$
		Negative to neutral word difference	Fixation time	−0.702	$p = 0.0187$
			Fixation latency	0.698	$p = 0.0207$

Note: Each feature was selected in at least 80% of LOOCV folds. Effect sizes were calculated according to Cohen's D criteria for the entire sample. Significance (p values) calculated using Welch's t test for the entire sample. RD task missed feedback features were missing and replaced for some participants, hence the effects of depressive symptoms were not shown.

and a high-resolution colour camera, together with a desktop computer. The accessibility and low cost of these components could make it easier to conduct replication or validation studies and could aid in translating results from research to other settings. In addition, the methods in our study could be more sensitive to first-episode and more mild depression, where structural brain abnormalities may not be present (e.g. Schmaal et al., 2016). It is possible that the studied methods could be more suitable for screening depression in primary healthcare settings – for example in general practices or community hospitals. MRI, on the other hand, could then be used in larger hospitals and specialised clinics for determining best courses of treatment for severe cases.

An important technical limitation in our study was participant exclusion due to face-tracking or eye-tracking problems. Face-tracking problems occurred for 12 participants and eye-tracking faults were present for another 11. In future studies face-tracking could be improved by increasing resolution of visual recordings, and by asking participants to wear a collar to make the jawline easily discernible during automated face-tracking analysis. Extra lighting focused on the participant's faces during the assessment could also be added. To reduce eye-tracking faults, additional methods to restrict participant movement may be explored, together with the application of eye-tracking devices with higher resolution.

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press

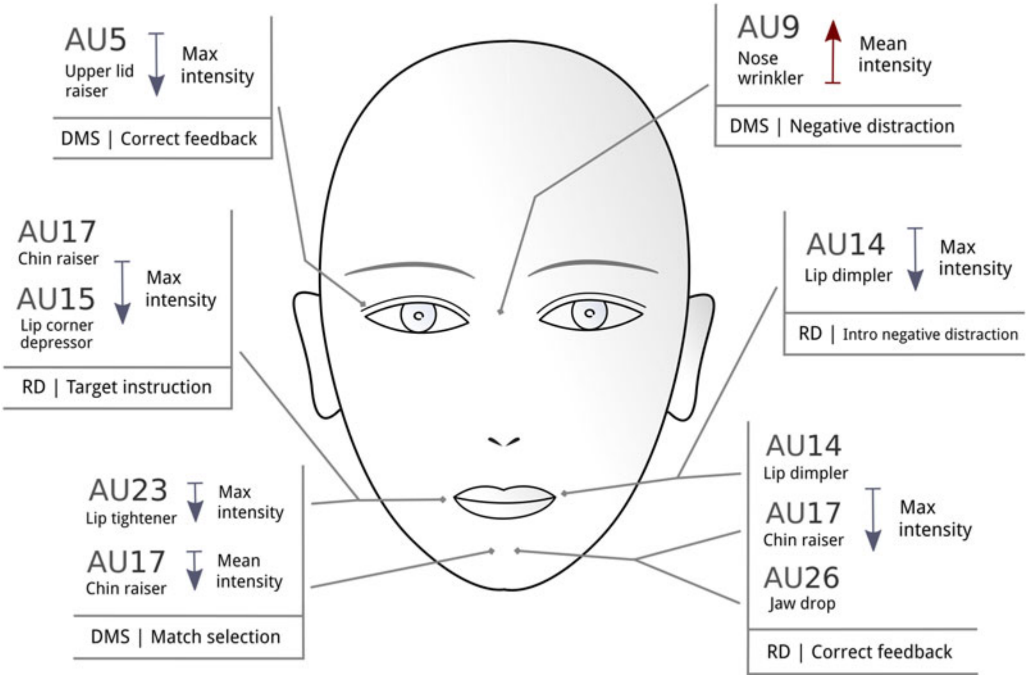

Fig. 2. Identified consensus face movement features which characterise symptomatic participants (excluding those related to the RD *missed-target feedback*). Each feature was selected in at least 80% of LOOCV folds.

Experiment participants

A general limitation of our study was the participant sample. We assessed largely young and non-clinical participants, recruited from a student population (Table 1). No clinical diagnostic information was used. We thus do not yet have evidence that the results are directly translatable to clinical populations or to participants with severe and enduring unipolar or bipolar illnesses. Further investigations should apply similar methods to investigate participants with formal diagnoses and other age groups, as well as larger sample sizes.

Control and symptomatic groups in our study were largely similar with regard to demographic characteristics, but symptomatic participants were on average 3.8 years younger compared to controls (Table 1). No studies to date indicate age-related differences in face or eye movements in young adults. We opted to avoid additional correction for age in order to preserve any effects of interest as much as possible, and also because all participants were young and the difference in age between the groups was relatively small. Future studies could investigate if face and eye movements may be differentially affected by depression in younger or older age.

Depressive symptom features

Face-tracking features

Symptomatic participants displayed *reduced* intensities of mouth or eyelid movements when selecting match pattern at the DMS task, when receiving target instruction at the RD task, and when receiving correct feedback at both the DMS and RD tasks (Table 2, Fig. 2). Cross-validation *consensus* features included maximal intensities of upper eyelid raiser (AU5), lip dimpler (AU14), lip corner depressor (AU15), chin raiser (AU17), and lip tightener (AU23). Reduced facial action intensities could be indicative of lower concentration on the task, although this was not reflected in reaction times and accuracies. To the best of our knowledge, no study to date has analysed face movements during cognitive task performance in depression – although patterns of reduced facial activity during clinical interviews have been reported (Cohn et al., 2009; Gaebel & Wölwer, 1992, 2004). Further work will be needed to validate our results in larger samples.

With regard to affective distractions – symptomatic participants displayed increased nose wrinkling (AU9) at the negative distraction stage of the DMS task and decreased maximal intensity of lip dimpling (AU14) at the introductory negative distraction stage of the RD task. This supports the proposition that facial reactions to negative affective material are altered in depression, although the precise pattern of changes needs to be confirmed in the future. The previous literature is inconsistent: some reports indicate *reduced* facial movements during sad mental imagery (Gehricke & Shapiro, 2000, 2001), and when viewing negative affective pictures or clips (Renneberg et al., 2005; Schneider et al., 1990; Wexler, Levenson, Warrenburg, & Price, 1994). Others, on the other hand, reported *increased* facial muscle activities (mouth and eyebrow) or increased facial movement frequencies in response to sad mental imagery (Greden, Genero, Price, Feinberg, & Levine, 1986; Schwartz, Fair, Salt, Mandel, & Klerman, 1976a, 1976b), depression-related thoughts (Teasdale & Bancroft, 1977; Teasdale & Rezin, 1978), or negative affective pictures (Sloan et al., 1997). It is possible that different contexts and different types of affective material are related to different effects in depression, and future studies could clarify which changes are related to which stimuli.

Eye-tracking features

All of the identified *eye-tracking* consensus features in our study were related to distraction words (Table 2). Previous literature indicates that depressed participants fixate more and for longer

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press

on negative material and less on positive material (reviews in Armstrong & Olatunji, 2012; Carvalho et al., 2015). In our study, symptomatic participants fixated fewer times on *positive* words during the DMS distraction stage and longer on *all* distraction words during the DMS selection stage, but no specific negative bias was observed. Lack of a negative bias effect could have been because the participants were a non-clinical group and because the distraction words in our study may have been less salient than affective material in other studies. Nonetheless, our results indicate that verbal affective distractions can be useful for detecting elevated symptoms of depression.

Conclusion

Our proof-of-concept study indicates that elevated symptoms of depression can in principle be predicted using face and eye movement tracking during cognitive task performance. Symptomatic participants were identified mainly by reduced intensities of mouth or eye movements during different stages of the cognitive tasks, as well as differences in eye fixations on verbal distraction stimuli. Future work will be needed to investigate larger samples and clinical participants, to improve the technical setup and reduce participant drop-out rates, and to define which specific depression symptoms may be related to which changes in the face and eye movements.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291720003608.

Acknowledgements. The current work was funded by grants EP/F500385/1 and BB/F529254/1 for the University of Edinburgh, School of Informatics, Doctoral Training Centre in Neuroinformatics and Computational Neuroscience from the UK Engineering and Physical Sciences Research Council (EPSRC), the UK Biotechnology and Biological Sciences Research Council (BBSRC), and the UK Medical Research Council (MRC).

Conflict of interest. JDS previously received funding via an honorarium associated with a lecture from Wyeth, and research funding from Indivior. None of these funding sources is connected to the present study. No potential conflicts of interest are reported for other authors.

References

Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. (2013). *Eye movement analysis for depression detection*. IEEE International Conference on Image Processing, Melbourne, VIC, 2013, pp. 4220–4224. https://doi.org/10.1109/ICIP.2013.6738869.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review*, *32*(8), 704–723. https://doi.org/10.1016/j.cpr.2012.09.004.

Baltrusaitis, T., Mahmoud, M., & Robinson, P. (2015). *Cross-dataset learning and person-specific normalisation for automatic Action Unit detection*. 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, 2015, pp. 1–6. https://doi.org/10.1109/FG.2015.7284869.

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

Bright, P., Jaldow, E., & Kopelman, M. D. (2002). The National Adult Reading Test as a measure of premorbid intelligence: A comparison with estimates derived from demographic variables. *Journal of the International Neuropsychological Society: JINS*, *8*(6), 847–854.

Carvalho, N., Laurent, E., Noiret, N., Chopard, G., Haffen, E., Bennabi, D., & Vandel, P. (2015). Eye movement in unipolar and bipolar depression: A systematic review of the literature. *Frontiers in Psychology*, *6*, 1809. https://doi.org/10.3389/fpsyg.2015.01809.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., … De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, 2009, pp. 1–7. https://doi.org/10.1109/ACII.2009.5349358.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018.

Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry: Official Journal of the World Psychiatric Association* (*WPA*), *13*(1), 28–35. https://doi.org/10.1002/wps.20087.

Ekman, P. E., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system (FACS)*. Salt Lake City, UT: Research Nexus.

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., … Yager, J. (2013). The initial field trials of DSM-5: New blooms and old thorns. *The American Journal of Psychiatry*, *170*(1), 1–5. https://doi.org/10.1176/appi.ajp.2012.12091189.

Gaebel, W., & Wölwer, W. (1992). Facial expression and emotional face recognition in schizophrenia and depression. *European Archives of Psychiatry and Clinical Neuroscience*, *242*(1), 46–52.

Gaebel, W., & Wölwer, W. (2004). Facial expressivity in the course of schizophrenia and depression. *European Archives of Psychiatry and Clinical Neuroscience*, *254*(5), 335–342. https://doi.org/10.1007/s00406-004-0510-5.

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037–1052. https://doi.org/10.1111/cns.13048.

Gehricke, J., & Shapiro, D. (2000). Reduced facial expression and social context in major depression: Discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research*, *95*(2), 157–167.

Gehricke, J., & Shapiro, D. (2001). Facial and autonomic activity in depression: Social context differences during imagery. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *41*(1), 53–64.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., & Rosenwald, D. P. (2014). Nonverbal Social Withdrawal in Depression: Evidence from manual and automatic analysis. *Image and Vision Computing*, *32*(10), 641–647. https://doi.org/10.1016/j.imavis.2013.12.007.

Greden, J. F., Genero, N., Price, H. L., Feinberg, M., & Levine, S. (1986). Facial electromyography in depression. Subgroup differences. *Archives of General Psychiatry*, *43*(3), 269–274.

Johnston, B. A., Steele, J. D., Tolomeo, S., Christmas, D., & Matthews, K. (2015). Structural MRI-based predictions in patients with Treatment-Refractory Depression (TRD). *PloS One*, *10*(7), e0132958. https://doi.org/10.1371/journal.pone.0132958.

Johnston, B. A., Tolomeo, S., Gradin, V., Christmas, D., Matthews, K., & Steele, J. D. (2015). Failure of hippocampal deactivation during loss events in treatment-resistant depression. *Brain: A Journal of Neurology*, *138*(Pt 9), 2766–2776. https://doi.org/10.1093/brain/awv177.

Kambeitz, J., Cabral, C., Sacchet, M. D., Gotlib, I. H., Zahn, R., Serpa, M. H., … Koutsouleris, N. (2017). Detecting neuroimaging biomarkers for depression: A meta-analysis of multivariate pattern recognition studies. *Biological Psychiatry*, *82*(5), 330–338. https://doi.org/10.1016/j.biopsych.2016.10.028.

McIntyre, R. S., Cha, D. S., Soczynska, J. K., Woldeyohannes, H. O., Gallaugher, L. A., Kudlow, P., … Baskaran, A. (2013). Cognitive deficits and functional outcomes in major depressive disorder: Determinants, substrates, and treatment interventions. *Depression and Anxiety*, *30*(6), 515–527. https://doi.org/10.1002/da.22063.

McLellan, T. M., Caldwell, J. A., & Lieberman, H. R. (2016). A review of caffeine's effects on cognitive, physical and occupational performance. *Neuroscience and Biobehavioral Reviews*, *71*, 294–312. https://doi.org/10.1016/j.neubiorev.2016.09.001.

Mwangi, B., Ebmeier, K. P., Matthews, K., & Steele, J. D. (2012). Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain: A Journal of Neurology*, *135* (Pt 5), 1508–1521. https://doi.org/10.1093/brain/aws084.

Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244. https://doi.org/10.1007/s12021-013-9204-3.

National Institute for Health and Care Excellence. (2009). NICE clinical guideline 90: Depression in adults: The treatment and management of depression in adults. Retrieved from guidance.nice.org.uk/cg90.

Pampouchidou, A., Simos, P. G., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M., & Tsiknakis, M. (2019). Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, *10*(4), 445–470. https://doi.org/10.1109/TAFFC.2017.2724035.

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*(3), 598–614. https://doi.org/10.3758/BRM.41.3.598.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1* (3), 385–401. https://doi.org/10.1177/014662167700100306.

Renneberg, B., Heyn, K., Gebhard, R., & Bachmann, S. (2005). Facial expression of emotions in borderline personality disorder and depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *36*(3), 183–196. https://doi.org/10.1016/j.jbtep.2005.05.002.

Rock, P. L., Roiser, J. P., Riedel, W. J., & Blackwell, A. D. (2014). Cognitive impairment in depression: A systematic review and meta-analysis. *Psychological Medicine*, *44*(10), 2029–2040. https://doi.org/10.1017/S0033291713002535.

Rosa, M. J., Portugal, L., Hahn, T., Fallgatter, A. J., Garrido, M. I., Shawe-Taylor, J., & Mourao-Miranda, J. (2015). Sparse network-based models for patient classification using fMRI. *NeuroImage*, *105*, 493–506. https://doi.org/10.1016/j.neuroimage.2014.11.021.

Rottenberg, J., Gross, J. J., & Gotlib, I. H. (2005). Emotion context insensitivity in major depressive disorder. *Journal of Abnormal Psychology*, *114*(4), 627–639. https://doi.org/10.1037/0021-843X.114.4.627.

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption--II. *Addiction* (*Abingdon, England*), *88*(6), 791–804.

Schmaal, L., Veltman, D. J., van Erp, T. G. M., Sämann, P. G., Frodl, T., Jahanshad, N., … … Hibar, D. P. (2016). Subcortical brain alterations in major depressive disorder: Findings from the ENIGMA Major Depressive Disorder working group. *Molecular Psychiatry*, *21*(6), 806–812. https://doi.org/10.1038/mp.2015.69.

Schneider, F., Heimann, H., Himer, W., Huss, D., Mattes, R., & Adam, B. (1990). Computer-based analysis of facial action in schizophrenic and depressed patients. *European Archives of Psychiatry and Clinical Neuroscience*, *240*(2), 67–76.

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., & Klerman, G. L. (1976a). Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science* (*New York, NY*), *192*(4238), 489–491.

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., & Klerman, G. L. (1976b). Facial expression and imagery in depression: An electromyographic study. *Psychosomatic Medicine*, *38*(5), 337–347.

Sloan, D. M., Strauss, M. E., Quirk, S. W., & Sajatovic, M. (1997). Subjective and expressive emotional responses in depression. *Journal of Affective Disorders*, *46*(2), 135–141.

Steele, J. D., & Paulus, M. P. (2019). Pragmatic neuroscience for clinical psychiatry. *The British Journal of Psychiatry: The Journal of Mental Science*, *215* (1), 404–408. https://doi.org/10.1192/bjp.2019.88.

Stratou, G., Scherer, S., Gratch, J., & Morency, L.-P. (2015). Automatic nonverbal behavior indicators of depression and PTSD: The effect of gender. *Journal on Multimodal User Interfaces*, *9*(1), 17–29. https://doi.org/10.1007/s12193-014-0161-4.

Teasdale, J. D., & Bancroft, J. (1977). Manipulation of thought content as a determinant of mood and corrugator electromyographic activity in depressed patients. *Journal of Abnormal Psychology*, *86*(3), 235–241.

Teasdale, J. D., & Rezin, V. (1978). Effect of thought-stopping on thoughts, mood and corrugator EMG in depressed patients. *Behaviour Research and Therapy*, *16*(2), 97–102.

Wang, L., Shen, X., Wu, Y., & Zhang, D. (2016). Coffee and caffeine consumption and depression: A meta-analysis of observational studies. *The Australian and New Zealand Journal of Psychiatry*, *50*(3), 228–242. https://doi.org/10.1177/0004867415603131.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x.

Wei, M., Qin, J., Yan, R., Li, H., Yao, Z., & Lu, Q. (2013). Identifying major depressive disorder using Hurst exponent of resting-state brain networks. *Psychiatry Research*, *214*(3), 306–312. https://doi.org/10.1016/j.pscychresns.2013.09.008.

Wexler, B. E., Levenson, L., Warrenburg, S., & Price, L. H. (1994). Decreased perceptual sensitivity to emotion-evoking stimuli in depression. *Psychiatry Research*, *51*(2), 127–138.

Zeng, L.-L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., … Hu, D. (2012). Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis. *Brain: A Journal of Neurology*, *135*(Pt 5), 1498–1507. https://doi.org/10.1093/brain/aws059.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006.

Zugal, S., & Pinggera, J. (2014). Low–cost Eye–trackers: Useful for information systems research? In L. Iliadis, M. Papazoglou, & K. Pohl (Eds.), *Advanced information systems engineering workshops* (Vol. *178*, pp. 159–170). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07869-4_14.

https://doi.org/10.1017/S0033291720003608 Published online by Cambridge University Press