Heliyon 6 (2020) e03572

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

An evaluation of the distribution properties, factor structure, and item response profile of an assessment of emotion recognition

Karen McKenzie [a,*,1], Aja Murray [b,1], Kara Murray [c], Michael O'Donnell [d], George C. Murray [c], Dale Metcalfe [a], Kristofor McCarty [a]

[a] Northumbria University, United Kingdom
[b] Edinburgh University, United Kingdom
[c] NHS Lothian, United Kingdom
[d] North Bristol NHS Trust, United Kingdom

ARTICLE INFO

Keywords:
Psychology
Assessment
Autism spectrum disorder
Emotion recognition
Item response profile
Intellectual disability

ABSTRACT

Many people with developmental disabilities, such as autism spectrum disorder and intellectual disability have emotion recognition (ER) difficulties compared with typically developing (TD) peers. Accurate assessment of the extent and nature of differences in ER requires an understanding of the response profiles to ER assessment stimuli. We analysed data from 504 TD individuals in response to an ER assessment in respect of distribution properties, factor structure, and item response profile. Eighteen emotion items discriminated better at lower levels of ER ability in TD participants. Neutral expressions were the hardest to interpret; surprise, anger, happy, and bored were easiest. The amount of contextual information in combination with the emotion being depicted also appeared to influence level of difficulty. Similar psychometric research is needed with people with developmental disabilities.

1. Introduction

Emotion recognition (ER) skills are considered to be important for socio-emotional development and functioning (Connolly et al., 2016), but some groups, such as people with developmental disabilities, have difficulties with ER. This includes individuals with an intellectual disability, who have been found to experience deficits with ER compared to their typically developing (TD) peers (Scotland et al., 2015). Similar difficulties in people with autism spectrum disorder have been found in many (see Uljarevic and Hamilton, 2013), but not all studies (see Harms et al., 2010).

A challenge for developing theoretical models that explain ER deficits in people with developmental disabilities, is that studies differ in the stimuli used, which can influence the way they are processed (Speer et al., 2007). Likewise, comparative research into those with and without ER difficulties requires knowledge about the properties of the ER assessment being used and whether they are equivalent in respect of their item response profiles across groups (Facon et al., 2011).

This highlights the need for research with large samples of TD individuals in order to understand the assessment properties of emotion stimuli. This is particularly important as many commonly used emotion stimuli are normed by asking respondents to choose what they think is the most appropriate descriptor from a restricted set of emotion labels, along with a rating of intensity (see Teh et al., 2018). Understanding item response properties can go beyond this, and help guide the selection of items to administer to people of differing levels of ability by indicating which items are the most and least difficult and which best differentiate between different levels of ability. Such research can also facilitate both the identification of factors that (differentially) affect performance in TD individuals and those with developmental disabilities and the development of effective, evidence-based interventions to improve ER skills. The introduction of whole school approaches and psychoeducational interventions to promote ER skills (Connolly et al., 2016; McKenzie et al., 2000) have also highlighted the need to use robust ER measures for assessment, teaching and evaluation purposes.

In this context, the present study aimed to identify the properties (difficulty and discrimination) of an assessment of ER when administered to a TD sample i.e. excluding those with conditions known to be associated with deficits in ER. The assessment was originally developed by McKenzie et al. (2001) and was subsequently updated with new and

* Corresponding author.
E-mail address: k.mckenzie@northumbria.ac.uk (K. McKenzie).
[1] Joint first author.

https://doi.org/10.1016/j.heliyon.2020.e03572
Received 20 February 2019; Received in revised form 1 October 2019; Accepted 9 March 2020

additional stimuli. which have been used and evaluated in a number of comparative studies of the ER of adults and children with and without a developmental disability (McKenzie et al., 2018; Murray et al., 2019; Scotland et al., 2016). The assessment stimuli also vary in the amount of contextual information available. This is important because research has indicated that the amount and type of contextual information can impact differentially on the accuracy of ER in individuals with and without a developmental disability (Barrett et al., 2011; McKenzie et al., 2001; Martin et al., 2019; Murray et al., 2019; Scotland et al., 2016; Teh et al., 2018).

As research with TD individuals suggests that the basic emotions (happy, sad, angry, afraid and disgust) are recognised universally, from a very young age (see Baron-Cohen et al., 2001), it is hypothesised that these emotions will be the easiest to identify and least discriminating. Similarly, as having contextual information that is relevant to the emotion being depicted has been found to increase the accuracy of ER (e.g. Barrett et al., 2011), it is hypothesised that the stimuli with contextual information relevant to the emotion being depicted will also be easier to identify and less discriminating in TD adults.

2. Materials and methods

2.1. Participants

Participants were included if they were adults who did not have a condition known to be linked with ER difficulties e.g. intellectual disability. This is was established by asking participants to note if they had any of the following conditions: learning difficulty, intellectual disability, autism spectrum disorder, physical disability, mental health problem, other condition (with a request to specify what this was) or none. Only those participants who selected 'none' as their response were included in the final analysis. After excluding any ineligible participants 504 people were included in the analysis (male = 131 (26%), female = 370 (73%), 'other' = 2 (0.4%), missing = 1), aged 18 to 97 (mean = 26.7 years, SD = 17.4). The majority of participants were from the United Kingdom (n = 323; 64%) or United States (n = 101; 20%). Most were employed (n = 239; 47%) or students (n = 140; 28%) with the remainder being retired or not in paid employment. In terms of educational level, only 32 (6%) had no qualification, while 156 (31%) had a degree, and 123 (24%) had a postgraduate qualification. The remaining participants had at least a qualification to standard grade or equivalent, i.e. broadly equivalent to school exams taken at age 16.

2.2. Design

Ethical approval for the study was obtained from Northumbria University, Department of Psychology ethics committee. The study utilised a cross-sectional design with data being gathered via an online assessment. All participants provided informed consent.

2.3. Procedure

Participants were recruited via information posted on a variety of online forums/social media sites and through word of mouth. Potential participants were provided with a link to the online study which provided detailed information about the research. If they consented to take part, by clicking on a 'consent' button, they gained access to the online assessment. Participants were initially asked to provide demographic information including age, gender, occupation, and highest level of education and whether the participant had any of the conditions outlined previously. The participants were then asked to complete the ER task as outlined below. The stimuli were presented one at a time and there was no time limit for responses. Once the participants entered their response and pressed the return button, the next image was displayed until the task was complete. Participant responses were anonymous and no feedback or compensation was provided.

2.3.1. Emotion recognition

This was based on an assessment developed by McKenzie et al. (2001) which was updated to include stimuli depicting six basic emotions of 'happy,' 'sad,' 'afraid,' 'angry,' 'surprised,' and 'disgusted.' It is recognised that there is some debate over which emotions constitute 'basic' emotions and the extent to which the concept of basic emotions is valid, however, in line with much previous ER research, those basic emotions initially identified by Eckman were chosen for inclusion (see Hutto et al., 2018). In addition, three others were included ('worried,' bored,' and 'neutral'), based on the suggestion of the study advisory group (see below), as these are commonly taught to individuals with developmental disabilities as part of interventions to improve ER. There is some debate about whether worry and boredom constitute cognitive states or emotions, however, as a number of researchers identify them as emotions (e.g., Hofmann et al., 2005; Jiang et al., 2014), they are included as such in the present study. 'Neutral' was also included as an indicator that the person was feeling 'OK.'

The assessment required participants to identify and label the nine emotions in line drawings, photos with limited context, and photos with context. The line drawings, which depicted only a face, were commissioned from an artist for the study. The artist was requested to depict each specific emotion using only a face, with no additional contextual information.

The photographs, all of which had a creative commons licence allowing their re-use, were sourced from Flickr (https://www.flickr.com/). A search was conducted using the name of each emotion. The first author then selected a number of examples of photographs from the results that were felt to best depict the specific emotion. These were then shared with a small group of advisors who had experience of working and researching in the field of emotion recognition. From this discussion, the stimulus that was agreed to best depict each emotion in question was chosen and all were subsequently piloted with a small non-clinical population sample to assess their face validity. The stimuli were then used and evaluated in a series of studies which included typically developing adults and children, and individuals with a developmental disability (McKenzie et al., 2018; Murray et al., 2019; Scotland et al., 2016).

The individuals in the final photographic stimuli varied in terms of age, gender, and race depicted. In terms of age, there were 15 older adults, 17 adults, and 12 children/adolescents. In terms of gender, 30 were women and 14 were men. The races depicted were: 28 White, 2 Black, 8 Asian, and 6 Hispanic people.

The participants were presented with three sets of individual pictures depicting the nine different emotions and asked to type the name of the emotion depicted in the picture by answering the question 'What is the person [or people if applicable] feeling?' The line drawing stimuli were 469 x 469 pixels. The photographs were scaled down so that they were clearly visible in full on all screens. They varied a little in size due to some being portrait and some landscape but all were between 300-500 pixels wide. All stimuli were displayed in the centre of each page.

Each set differed in the extent to which emotional cues were available, beginning with line drawings of faces through photographs of people with limited emotional context to photographs with emotional context. A stimulus was considered to have limited emotional context if it depicted some additional details, but these did not give situational clues about the emotion being depicted (e.g. a man smiling, but with no situational cues to indicate that he was happy). A stimulus was considered to have emotional context if the additional details gave situational cues that were congruent with the emotion being depicted (e.g. people looking happy at a wedding). Figure 1 illustrates the three levels of context for the emotion 'happy.' The original stimuli were in colour. The same nine emotions were depicted for each level of contextual information. Copies of the materials used in the study can be obtained from the first author.

Responses were scored using a computerised scoring system. This was originally populated with synonyms for each of the emotions being depicted, which were sourced from the thesaurus of an Apple computer

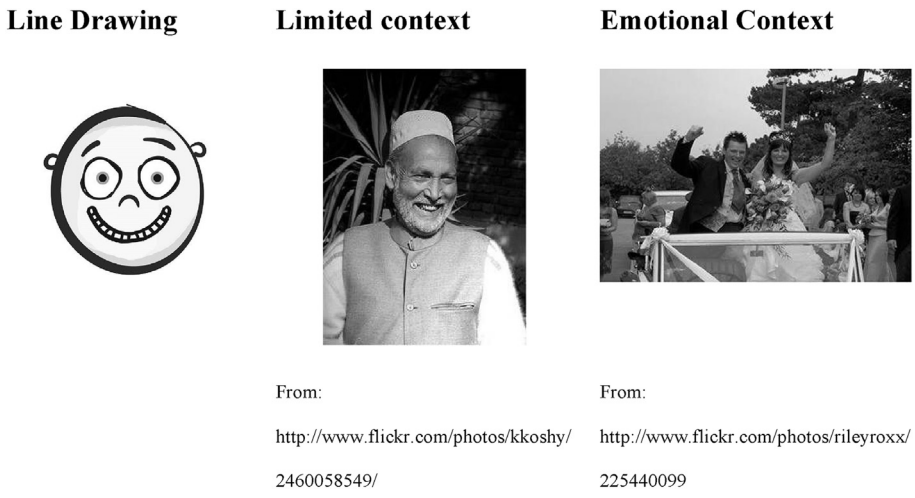

Figure 1. Examples of different levels of context for the emotion 'Happy'. Note: Photographs reproduced under creative commons licence Attribution 2.0 Generic (CC BY 2.0) https://creativecommons.org/licenses/by/2.0/.

e.g. 'joyful,' 'cheerful' were deemed acceptable responses for 'happy.' The databank of acceptable responses was added to as the responses were scored (see below). The computer was programmed to identify a response as correct, incorrect or unknown. 'Unknown' responses (e.g. due to spelling mistakes or words that were not in the original databank) were highlighted and subsequently coded by the research team as correct or incorrect. Any responses that were considered to be correct were added to the computer databank of correct responses, while incorrect responses, were coded as such by the programme.

2.4. Analysis strategy

2.4.1. Sample size

We calculated minimum sample size required based on a two parameter logistic IRT model (2PL model) and the total number of ER items. Based on the recommendations of Şahin and Anıl (2017), a minimum sample size of 250 was required. Şahin and Anıl (2017) studied the necessary minimum sample sizes to estimate the parameters of the 2PL model given different test lengths. They used correlations between parameters estimated in a large base sample of n = 6288 and parameters estimated in smaller subsamples of varying sizes and root mean squared difference (RMSD) as their criteria for determining necessary sample size. For a test length of 30 items, the minimum recommended sample size to achieve high parameter correlations and RMSD was 250.

2.4.2. Item selection

We describe a multi-step procedure by which we identified the 'best' set of items in terms of range difficulty and discrimination from the initial pool. Here accuracy, i.e., correctly identified or not, is being used as the measure of item difficulty. While some items were excluded at each stage, for example due to ceiling effects, such items may still provide useful information in assessment contexts and/or comparative research. The numbers in parentheses for each section heading indicate the number of items included in the pool at the beginning of that stage. We began by examining basic descriptive properties of the items (item distributions and correlations). This was followed by factor retention techniques to explore the dimensionality of the item set. We then used CFA to provide further diagnostic information, in particular, to identify potential violations of local independence (residual covariances). Finally, we used IRT to estimate the item properties themselves and associated test information function.

2.4.3. Missingness

At the item level, missingness ranged from 5% up to 22% for the emotion stimuli. For the preliminary analyses of proportion correct and item inter-correlations, factor retention methods, and confirmatory factor analysis (CFA), we used pairwise deletion. For the main psychometric, item response theory (IRT) analyses, missingness were dealt with via maximum likelihood estimation.

3. Results

3.1. Floor/ceiling effects (i = 27)

Our initial item pool included 27 items. Items were considered as showing very little variability in responses if the proportion correct was <5% or >95%. This criterion was chosen as a pragmatic cut-off with the aim of achieving a good balance between including items that had a good range of difficulty (including very easy and very difficult items) while excluding items that were likely to be relatively uninformative about performance level in the target population. None of the emotion stimuli showed proportion correct >95%, therefore all were retained for the next stage of analysis. Table 1 shows the number and percentage of participants correctly identifying each of the emotions in each condition. Tables 2 and 3 provide this information, stratified by gender and age respectively. As research suggests that older adults (commonly defined as those age 65 years and older) have greater difficulty with emotion recognition than younger adults (e.g. Abbruzzese et al., 2019; Goncalves et al., 2018; Isaacowitz et al., 2007; Franklin and Zebrowitz, 2017), age groups are those aged under 65 years and those aged 65 years and above in Table 3.

3.2. Correlations with other items (i = 27)

We then identified any items that had an extremely low correlation with other items. It is important to allow for some low to moderate correlations between some items because restricting the set of items to the most highly correlated risks selecting a set of highly similar items and thus restricting the breadth of content of the assessment. However, very low or negative correlations with other items may suggest problems with reliability. To account for the binary response format of items, we computed tetrachoric correlations among all the items (Olsson, 1979). We did not identify any items to be excluded on the basis of very low or negative correlations with other items.

3.3. Factor retention (i = 27)

To assess how many factors were optimal, we used parallel analysis with principal components analysis (PA-PCA) and the minimum average partial (MAP) test, and visual inspection of a scree plot. PA-PCA

Table 1. Number and percentage of participants correctly identifying each emotion under the three conditions.

Emotion Stimuli	Line drawing	Photos with little context	Photos with more context
	Number correct (percentage)		
Sad	399 (79)	274 (54)	205 (41)
Worried	228 (45)	261 (52)	165 (33)
Happy	402 (80)	392 (78)	351 (70)
Surprise	396 (78.5)	372 (74)	352 (70)
Disgust	295 (59)	317 (63)	321 (64)
Bored	51 (10)	277 (55)	338 (67)
Angry	416 (82.5)	372 (74)	290 (58)
Afraid	176 (35)	105 (21)	255 (51)
Neutral	170 (34)	92 (18)	106 (21)

Note. 'Number correct' refers to the number of the total sample who answered the item correctly; 'percentage' refers to the percentage of the total sample who answered the item correctly.

suggested the retention of three dimensions, while MAP and the scree plot (Figure 2) suggested the retention of one dimension. The scree plot in Figure 2 shows the eigenvalues (a measure of proportion of variance explained) for each successive principal component. Scree plots can be inspected for a point of inflection, with the number of dimensions to retain suggested to be the number of dimensions before that point. Given that the second eigenvalue only marginally exceeded the corresponding reference eigenvalue in the PA-PCA analysis, we judged retaining only one dimension to be the overall optimal solution. This suggested that neither the assessment of multiple emotions nor the use of multiple item types introduced substantial multi-dimensionality.

3.4. Confirmatory factor analysis (i = 27)

A single factor CFA was fit. Weighted least squares means and variances (WLSMV) estimation was used to account for the categorical response format of the items. Initially the model provided good fit to the data, with the root mean square error of approximation (RMSEA) < 0.05. However, the factor loadings for the items using line drawings were generally small (<0.30, often <0.10). Estimating the model without these items, the fit remained good (RMSEA <0.05) and the standardised loadings were now all >0.30 (mostly >0.40). Modification indices suggested that residual covariances between items measuring the same emotion should be estimated, indicating clustering of emotions together irrespective of context. We, therefore, included these parameters to avoid the inflation of factor loadings except between the 'afraid,' and 'disgust' item pairs where the residual covariance was minimal. The model is summarised in Figure 3. Item residual covariances are omitted for visual clarity. These were $r = .37$ for worried; $r = .53$ for surprised; $r = .54$ for bored; $r = .44$ for angry; $r = .52$ for OK and $r = .12$ for happy.

3.5. Item response theory analysis (i = 18)

We fit a two parameter logistic IRT model to the ER items (excluding the line drawings), accounting for residual covariances between items measuring the same emotion using specific factors orthogonal to one another and to the general factor. Identification of specific factors was achieved by fixing their variances to 1 and their indicator loadings equal to one another. We did not include specific factors for the 'afraid' and 'happy' item pairs because these did not appear to be causing substantive violations of local dependence. See Table 4 for parameter estimates and Figure 4 for the test information curve. The test information curve in Figure 4 shows how the test information (which is inversely related to measurement error) varies across different levels of latent trait values.

Table 2. Number and percentage of participants, stratified by gender, correctly identifying each emotion under the three conditions.

Emotion Stimuli	Line drawing	Photos with little context	Photos with more context	Line drawing	Photos with little context	Photos with more context
	Females			Males		
	Percentage correct					
Sad	90.8	66.6	52.1	91.1	57.4	52.4
Worried	54	61.8	43.9	46.4	60.7	35.9
Happy	92.3	92.7	88.3	89.3	90.7	91.3
Surprise	88.9	88	90.3	93.7	85	86.4
Disgust	67.6	76.6	83	66.1	68.2	76.7
Bored	12.3	67.1	87.5	8	60.4	80.6
Angry	94.8	89.2	75.9	95.5	83	67
Afraid	39.8	23.1	68.6	41.1	29.2	53.4
Neutral	40.6	20.5	28	34.8	25.2	24.3

Table 3. Number and percentage of participants, stratified by age, correctly identifying each emotion under the three conditions.

Emotion Stimuli	Line drawing	Photos with little context	Photos with more context	Line drawing	Photos with little context	Photos with more context
	Under 65 years			65 years and above		
	Percentage correct					
Sad	92	67	53	78	19	33
Worried	55	63	44	11	37	15
Happy	92	93	89	93	81	89
Surprise	92	88	91	67	74	67
Disgust	68	76	83	56	55	56
Bored	12	67	88	11	37	59
Angry	95	89	73	96	70	78
Afraid	41	25	66	22	22	48
Neutral	38	23	29	48	7	3

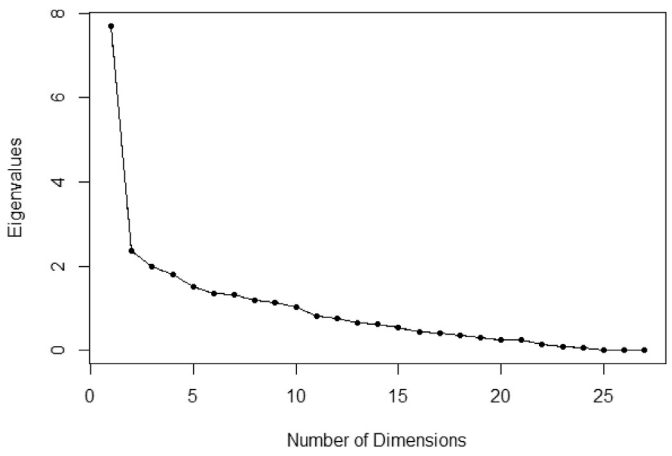

Figure 2. Scree plot.

Latent trait values are in standard deviation units. The maximum test information (the largest peak in Figure 4) was 8.59, equivalent to a classical test theory reliability of 0.88. Test information remained above 3.33 (equivalent to a classical test theory reliability of 0.70) between the latent trait values of -3.17 and 1.62 on the standard deviation scale.

Overall, the test was able to measure low levels of ER ability well, but its performance was poorer for discriminating ER levels in individuals with higher levels of ER ability. The difficulty parameters showed that there was no difference in difficulty between the photos with and without context; however, there were differences according to the emotions displayed. The items 'surprise,' 'happiness,' the item with more context showing 'boredom,' and the item showing 'anger' without context were the easiest. Items showing neutral expressions were most difficult in this TD sample.

4. Discussion

Researchers have emphasised the importance of identifying the item response profiles of participant groups when undertaking comparative assessments that involve individuals with developmental disabilities (Facon et al., 2011). Without knowing if the response profiles are equivalent, it is difficult to determine the nature of any differences found between groups. To this end, the study aimed to examine the item response properties of ER assessment stimuli used with a large sample of TD individuals.

It was hypothesised that the emotion stimuli with more contextual information would be less discriminating in TD adults, based on previous research (Barrett et al., 2011; McKenzie et al., 2001; Murray et al., 2019; Scotland et al., 2016). The study, however, found that some of the line drawing emotion stimuli had low factor loadings on the single factor model. Others (happy, sad, angry and surprised), while not reaching the exclusion threshold of >95% correct, showed percentage correct levels of between 78.5 and 82.5%. This indicates ceiling effects i.e. the items were too easy for the study sample. An obvious disadvantage of such ceiling effects is that the use of the stimuli would distort any group by task interactions because the tasks are not measuring the full range of abilities of the participants. Items with very high accuracy scores can, however, provide useful information as 'screening' items, in both educational contexts and comparative research, indicating that a response is out of the ordinary and may merit further investigation. In such cases, an incorrect response, when most other TD people with both high and low levels of ability on a task respond correctly, may indicate that the individual has difficulty with underlying basic abilities, such as linking abstract verbal concepts and visual stimuli. In addition, such items, while not discriminating in the TD population, may provide useful information when used with people with developmental disabilities.

Line drawings and symbol-based systems are commonly used in educational and other settings in an attempt to support the comprehension and communication of individuals with developmental disabilities (see Poncelas and Murphy, 2007). The results from the present study indicate a need to establish the impact of the inclusion of such materials on the item response profiles of those being assessed. It may be that materials that assist individuals with developmental disabilities may result in a ceiling effect for TD individuals, potentially masking the true extent of any differences in ER between those with and without developmental disabilities. Similarly, using ER stimuli that have been standardised with TD individuals may result in floor effects if used with people with developmental disabilities.

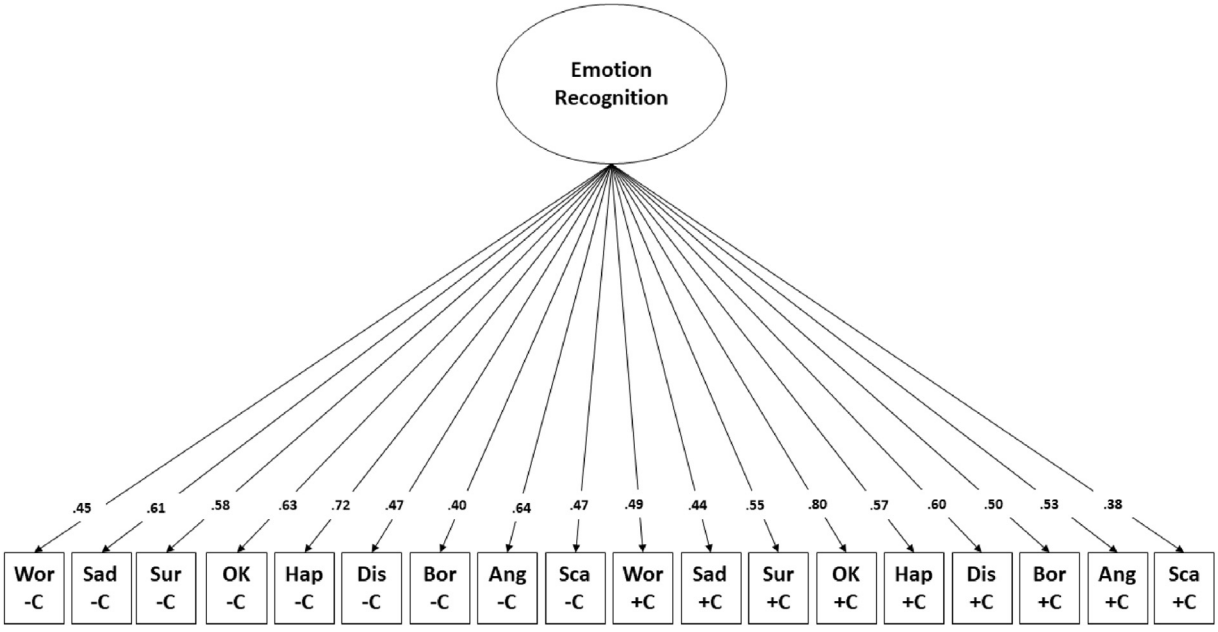

Figure 3. CFA model results. Note. Wor = worried, Sur = surprised; Hap = happy; Dis = disgust; Ang = Anger; -C = with limited contextual information; +C = with contextual information.

Table 4. Item response theory (2PL)[a] model parameter estimates.

Item	Discrimination general (standardised factor loading)	Location
LC[b] Worried	1.161 (0.467)	-0.708
LC Sad	1.293 (0.580)	-0.791
LC Surprised	1.967 (0.602)	-3.694
LC Neutral	1.931 (0.621)	2.624
LC Happy	2.072 (0.752)	-3.810
LC Disgust	0.832 (0.417)	-1.240
LC bored	1.173 (0.403)	-1.148
LC Angry	2.245 (0.680)	-3.770
LC Afraid	0.895 (0.442)	1.283
WC[c] Worried	1.181 (0.473)	0.483
WC Surprised	1.864 (0.582)	-3.912
WC Disgust	1.227 (0.560)	-1.861
WC Bored	1.528 (0.497)	-3.269
WC Angry	1.511 (0.530)	-1.779
WC Afraid	0.721 (0.370)	-0.670
WC Sad	0.853 (0.425)	-0.094
WC Neutral	3.105 (0.787)	2.304
WC Happy	1.466 (0.629)	-2.747

Note. 'Discrimination general' is the discrimination parameter for the general emotion recognition factor. For clarity, discrimination values for the specific factors are not shown.

[a] 2PL: two parameter logistic.

[b] LC: limited contextual information.

[c] WC: with contextual information.

The results from the item response theory analysis indicated that the 18 remaining emotion stimuli could discriminate at the lower end of ER ability but more poorly at the higher end. Individually, the items had moderate discrimination values, but the assessment as a whole appeared to provide an internally consistent measure of ER ability. The test information curve suggested that, as a set, the items could provide a reliable assessment of ER for a good range of abilities. The next step is to determine if the items discriminate to the same extent when used with those with developmental disabilities.

It was also hypothesised that the basic emotions (happy, sad, angry, afraid and disgust) would be least discriminating in our sample. The hypothesis was partly supported, with the results showing that the emotions displayed did vary in difficulty, but that the amount of contextual information available influenced this. The most difficult items showed neutral expressions, while the easiest were items showing 'surprise,' 'happiness,' 'boredom' with context, and 'anger' without context. A key question for future research will be whether the patterns of difficulty are similar in individuals with developmental disabilities. Such information will help researchers to better interpret the results of comparative studies, which in turn will assist in informing theories of ER. For example, some research has suggested that the recognition of 'fear' is more difficult for people with autism spectrum disorders as compared with 'happiness' (Uljarevic and Hamilton, 2013) and that people with an intellectual disability have a particular impairment on neutral expressions (Scotland et al., 2015), it will be valuable to directly compare ordering of item difficulties in the same item set.

Overall, our results highlight that it cannot be assumed that ER assessments will be interchangeable in different contexts without modification. For example, among the set of items presented in the current study, a teacher and comparative researcher would likely be interested in the responses to different items. The emotion stimuli that showed ceiling effects here may be of particular interest to a teacher, for example to rule out basic perceptual and cognitive deficits or start an individual on an easier set of ER stimuli to preserve motivation. A researcher interested in assessing group by task interactions to quantify ER difficulties at the group level would, however, likely be more interested in responses to items without ceiling effects in order to take account of any distortions of ER difficulties among those with developmental disabilities.

Our study did have some limitations. While the sample was relatively large and represented a population with a wide age range, from different occupational and educational backgrounds, and countries, most were from the UK or the US, the majority were female, and all were adults (with a mean age of 26.7 years). All of these factors could potentially influence the results, although in relation to age, research suggests that ability on ER tasks, is fairly consistent across most of adulthood, but can differ between younger and older adults (often defined as aged 65 upwards) for some emotions, with older adults having lower accuracy (e.g. Abbruzzese et al., 2019; Goncalves et al., 2018; Isaacowitz et al., 2007; Franklin and Zebrowitz, 2017). In respect of the influence of the gender of participants on responses, a recent meta-analysis by Goncalves et al. (2018) found gender differences on only two emotions, fear and disgust. Females performed better on the fear stimuli and worst on the disgust stimuli when compared to men. Research by de Souza et al. (2018), found no gender differences in responses to the Facial Emotion Recognition Test in Brazilian and French participants and no differences between participants from the different countries.

A meta-analysis by van Hemert et al. (2007) suggests that the average effect size of cultural differences in ER is small when corrected for artefacts, but some cultural differences do remain which are influenced by factors such as whether the emotion is positive or negative, religiosity, and political systems of the countries.

A further limitation was that the photo stimuli were not matched for factors such as age and gender, and the size and presentation varied slightly depending on whether the image was portrait or landscape layout. All of these factors may have influenced the results to some extent, although research has consistently found that individuals are not any better at recognising emotions from stimuli depicting people of the same age as themselves (see Vetter et al., 2018). Teh et al. (2018) note that the role of gender in emotion recognition has not been addressed in many commonly used emotion stimuli and this is an area for future research, along with further exploration of the influence of other aspects of the stimuli.

The main aim of the study was, however, to explore item response properties of the stimuli, with a particular emphasis on level of contextual information. Given this, it would have been very difficult to match the images, as contextual information relevant to a particular emotion can vary considerably.

A further limitation was that participants provided self-report information about any condition, such as intellectual disability, which might have impacted on their ER scores, rather than this being directly assessed. All responses were, however, anonymous and it seems unlikely that participants would not self-report a condition because of concerns about anonymity or confidentiality. A recent systematic review suggests that self-report measures about conditions such as anxiety and depression are generally completed equally as reliably in digital, online formats as compared with pen and paper versions (Alfonsson et al., 2014).

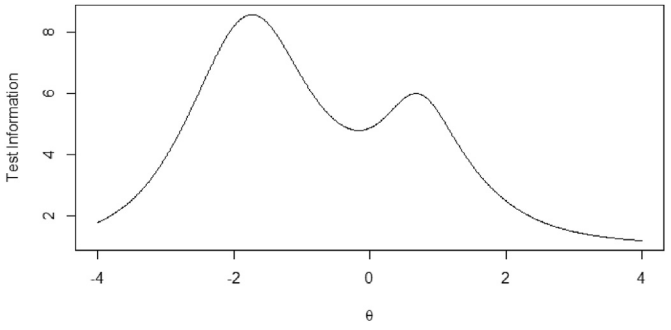

Figure 4. The test information curve.

It would also have been preferable to have multiple exemplars of each emotion in each condition in order to further determine the most robust examples in terms of their psychometric properties. This approach was precluded by the time and resources available to the project. In addition, administering very large numbers of items creates participant burden and potentially reduces the validity of their responses due to fatigue, boredom, or frustration. In this study, we were most concerned with ensuring maximally valid responses.

Finally, the study focused on only one ER assessment. Similar research is needed with other commonly used ER assessments, particularly as Teh et al. (2018) highlight that the way in which many commonly used emotion stimuli databases have been normed is to ask participants to choose the label, from options provided, that best describes the emotion and rate its intensity, rather than explore the item response properties of the stimuli.

4.1. Conclusion

We examined the distributional properties, factor structure, and item response profile of an ER assessment. Eighteen items were retained and their properties in terms of discrimination and difficulty were examined. The most difficult emotion to identify depicted a neutral expression and the easiest was 'surprise' when portrayed with context. The use of items with known and favourable psychometric properties will be important in advancing the assessment and understanding of ER difficulties, particularly in people with developmental disabilities. Without this information, it is difficult to know if differences are due to the nature of the stimuli used, such as amount of available contextual information and emotion being portrayed, rather than characteristics of the groups of people being assessed. This, in turn, makes it more difficult to develop robust theories and interventions to facilitate ER in people with developmental disabilities.

Data availability

Data are available from the second author.

Declarations

Author contribution statement

K. McKenzie: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

A. Murray: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

M. O'Donnell, K. Murray: Performed the experiments; Contributed reagents, materials, analysis tools or data.

G. Murray: Conceived and designed the experiments; Analyzed and interpreted the data.

D. Metcalfe, K. McCarty: Performed the experiments; Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Abbruzzese, L., Magnani, N., Robertson, I.H., Mancuso, M., 2019. Age and gender differences in emotion recognition. Front. Psychol. 10, 2371.

Alfonsson, S., Maathz, P., Hursti, T., 2014. Interformat reliability of digital psychiatric self-report questionnaires: a systematic review. J. Med. Internet Res. 16 (12), e268.

Barrett, L.F., Mesquita, B., Gendron, M., 2011. Context in emotion perception. Curr. Dir. Psychol. Sci. 20, 286–290.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., 2001. The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. J. Child Psychol. Psychiatry Allied Discip. 42 (2), 241–251.

Connolly, P., Miller, S., Mooney, J., Sloan, S., Hanratty, J., 2016. Universal school-based programmes for improving social and emotional outcomes in children aged 3-11 years: a systematic review and meta-analysis. In: The Campbell Collaboration Systematic Reviews. Retrieved from. http://www.campbellcollaboration.org/lib/project/369/.

de Souza, L.C., Bertoux, M., Vaz de Faria, A.R., Corgosinho, L.T.S., 2018. The effects of gender, age, schooling, and cultural background on the identification of facial emotions: a transcultural study. Int. Psychogeriatr. 30 (12), 1861–1870.

Facon, B., Magis, D., Belmont, J.M., 2011. Beyond matching on the mean in developmental disabilities research. Res. Dev. Disabil. 32, 2134–2147.

Franklin, R.G., Zebrowitz, L.A., 2017. Age differences in emotion recognition: task demands or perceptual dedifferentiation? Exp. Aging Res. 43 (5), 453–466.

Goncalves, A.R., Fernandes, C., Pasion, R., Ferreira-Santos, F., Barbosa, F., Marques-Teixeira, J., 2018. Effects of age on the identification of emotions in facial expressions: a metaanalysis. Peer J 6, e5278.

Harms, M.B., Martin, A., Wallace, G.L., 2010. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. Neuropsychol. Rev. 20 (3), 290–322.

Hofmann, S.G., Moscovitch, D.A., Litz, B.T., Kim, H.-J., Davis, L.L., Pizzagalli, D.A., 2005. The worried mind: autonomic and prefrontal activation during worrying. Emotion 5, 464–475.

Hutto, D.D., Robertson, I., Kirchhoff, M.D., 2018. A new, better BET: rescuing and revising Basic Emotion Theory. Front. Psychol. 9, 1217.

Isaacowitz, D.M., Löckenhoff, C.E., Lane, R.D., Wright, R., Sechrest, L., Riedel, R., Costa, P.T., 2007. Age differences in recognition of emotion in lexical stimuli and facial expressions. Psychol. Aging 22 (1), 147–159.

Jiang, Y., Hu, Y., Wang, Y., Zhou, N., Zhu, L., Wang, K., 2014. Empathy and emotion recognition in patients with idiopathic generalized epilepsy. Epilepsy Behav. 37, 139–144.

McKenzie, K., Matheson, E., McKaskie, K., Hamilton, L., Murray, G.C., 2000. The impact of group training on emotion recognition in individuals with a learning disability. Br. J. Learn. Disabil. 28, 1–6.

McKenzie, K., Matheson, E., McKaskie, K., Hamilton, L., Murray, G.C., 2001. A picture of happiness: emotion recognition in individuals with a learning disability. Learn. Disabil. Pract. 4 (1), 26–29.

McKenzie, K., Murray, A.L., Wilkinson, A., Murray, G.C., Metcalfe, D., O'Donnell, M., McCarty, K., 1 January 2018. The relations between processing style, autistic-like traits and emotion recognition in individuals with and without Autism Spectrum Disorder. Pers. Indiv. Differ. 120, 1–6. Early view.

Martin, R., McKenzie, K., Metcalfe, D., Pollet, T., McCarty, K., 2019. A preliminary investigation into the relationship between empathy, autistic like traits and emotion recognition. Pers. Indiv. Differ. 137, 12–16.

Murray, G.C., McKenzie, K., Murray, A.L., Whelan, K., Cossar, J., Murray, J., Scotland, J., January 2019. The impact of contextual information on the emotion recognition of children with an intellectual disability. J. Appl. Res. Intellect. Disabil. 32 (1), 152–158. Early view.

Olsson, U., 1979. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika 44, 443–460.

Poncelas, A., Murphy, G., 2007. Accessible information for people with intellectual disabilities: do symbols really help? J. Appl. Res. Intellect. Disabil. 20, 466–474.

Şahin, A., Anıl, D., 2017. The effects of test length and sample size on item parameters in Item Response Theory. Educ. Sci. Theor. Pract. 17, 321–335.

Scotland, J., Cossar, J., McKenzie, K., 2015. The ability of adults with an intellectual disability to recognise facial expressions of emotion in comparison with typically developing individuals: a systematic review. Res. Dev. Disabil. 41–42, 22–39.

Scotland, J., McKenzie, K., Cossar, J., Murray, A.L., Michie, A., 2016. Recognition of facial expressions of emotion by adults with intellectual disability: is there evidence for the emotion specificity hypothesis? Res. Dev. Disabil. 48, 69–78.

Speer, L.L., Cook, A.E., McMahon, W.M., Clark, E., 2007. Face processing in children with autism. Autism 11 (3), 265–277.

Teh, E., Yap, M.J., Liow, S.J.R., 2018. PiSCES: pictures with social context and emotional scenes with norms for emotional valence, intensity, and social engagement. Behav. Res. Methods 50 (5), 1793–1805.

Uljarevic, M., Hamilton, A., 2013. Recognition of emotions in autism: a formal meta-analysis. J. Autism Dev. Disord. 43, 1517–1526.

van Hemert, D.A., Poortinga, Y.H., van de Vijver, F.J.R., 2007. Emotion and culture: a meta-analysis. Cognit. Emot. 21, 913–943.

Vetter, N.C., Drauschke, M., Thieme, J., Altgassen, M., 2018. Adolescent basic facial emotion recognition is not influenced by puberty or own-age bias. Front. Psychol. 9. ISSN: 1664-1078.