

The Ability to Understand the Experience of Other People:
Development and Validation of the Emotion Recognition Scales

Murray J. Dyck
School of Psychology
Griffith University

Running head: EMOTION RECOGNITION SCALES

Word count: 6,402

Correspondence to:

Murray J. Dyck
School of Psychology
Griffith University
Gold Coast, Queensland, Australia, 4222

Tel: 61 7 5552 8251

Fax: 61 7 5552 8291

Email: m.dyck@griffith.edu.au

Abstract

The Emotion Recognition Scales (ERS) were developed to assess the ability to recognise facial and vocal expressions of common emotions, to understand the meaning of emotion terms, to understand relationships between emotions and the experiences that elicit them, and to use reasoning skills and knowledge of emotion-event relationships to resolve apparently incongruous emotional outcomes. The ERS were needed to supplement the set of objective assessment tools available to measure hypothesised deficits in social cognitive abilities in several populations. The ERS have been administered to a large representative sample of children and children with a range of disorders, including autism, intellectual disability, communication, motor skills, and attention disorders, deafness and blindness. The aim of this article is to describe the development of the ERS, summarise evidence on the reliability and validity of the ERS, and provide age norms for each of the ERS subtests.

The impetus for constructing the Emotion Recognition Scales (ERS) was a need to assess “empathic ability,” broadly defined as the ability to understand the experience of other people. The Swedish paediatrician Christopher Gillberg had proposed a new class of disorders that he called “empathy disorders” that referred to the inability “to conceptualise other people’s inner worlds and to reflect on their thoughts and feelings” (Gillberg, 1992, p. 835). In Gillberg’s view, Autistic Disorder and Asperger’s Disorder were the best exemplars of empathy disorders, but a low “empathy quotient” was also thought to characterise people with deficits in attention, motor control, and perception (e.g., Attention Deficit Hyperactivity Disorder, Developmental Coordination Disorder), Tourette’s Disorder, Intellectual Disability, Anorexia Nervosa, and Obsessive Compulsive Personality Disorder. Unfortunately, in the mid-1990s there was no reliable and valid way of measuring empathic ability, scant information on the normal development of empathic ability, and hence no way of assessing people’s ability to understand the experience of other people.

Gillberg (1992, 1993) equated empathic ability with the acquisition of a theory of mind, the realisation by young children that other people have minds distinct from their own and the recognition that knowledge of the mental states of other people is important to understanding the behaviour of other people (Wimmer & Perner, 1983). Consistent with Gillberg’s proposal, theory of mind deficits are pronounced among children with Autistic Disorder (Baron-Cohen, Leslie, & Frith, 1985), but are also evident in children with a number of disorders usually first evident in childhood, including Asperger’s Disorder, Attention Deficit Hyperactivity Disorder, and Intellectual Disability (Dyck, Ferguson, & Shochet, 2001). However, people with these disorders also have many other deficits, and so the presence of theory of mind

deficits in a range of disorders does not indicate if these deficits are more or less pronounced than other deficits associated with these disorders.

Theory of mind tasks are not well suited to assessing if empathic ability deficits are greater than other deficits in people with an empathy disorder, for three reasons. First, children normally acquire a theory of mind by their fourth year even though the ability “to conceptualise other people’s inner worlds” continues to develop well beyond this age. Theory of mind measures were developed to assess whether or not a given child has acquired a theory of mind. This means that the theory of mind construct and theory of mind measures are useful in understanding and assessing the developmental origins of an important component of empathic ability, for defining and measuring severe empathic ability deficits and for the early identification of empathic ability deficits, but they are not useful in understanding and assessing the ongoing development of broad social cognitive abilities. Second, acquisition of a theory of mind is represented as a qualitative and categorical change in cognitive ability, whereas the ability to understand others, from which an empathy quotient might be derived (Gillberg, 1992), implies a continuous distribution of individual differences in ability. Third, although acquisition of a theory of mind can be regarded as a prerequisite for understanding the experience of others, once it has been acquired, the ability to understand other people will subsequently depend on other specific abilities.

The ERS were developed to overcome these limitations and to measure other important components of the empathic ability construct. As their name implies, the ERS are designed to measure a person’s ability to understand the emotional experience of another person. This focus on understanding emotional experience was intended to reflect what is understood by the term empathy in traditional clinical

usage (Beck, Rush, Shaw, & Emery, 1979; Kohut, 1971; Rogers, 1951) and to reflect how developmental psychologists differentiate the theory of mind construct from other abilities that are essential for understanding other people.

Studying the “child-as-psychologist,” Dunn (1995) distinguished between the child’s need to understand the emotions of other people and the need to understand the minds of other people. Following Flavell (1992), Eisenberg, Murphy and Shepard (1996, p. 74) argued that in order to “understand the origins of perspective taking, it is necessary to review literature on children’s understanding of emotion and rudimentary mental constructs.” Eisenberg et al. distinguished two components of the ability to understand emotions, namely, the ability to decode and label emotions based on perceptual cues and the ability to use situational cues to make inferences about others’ emotions. These are the primary abilities that the ERS were designed to measure.

This article provides information on the construction of the ERS as well as comprehensive psychometric information and age-norms for the main subtests.

Test Construction

Broadly defined empathic ability was conceived as analogous to broadly defined general intelligence, that is, as a higher-order ability that emerges from a larger set of primary abilities, each of which has some commonality with each other primary ability. Also, as a set of abilities, the so-called empathic abilities would not be independent of the abilities that are used to operationally define intelligence but could, in fact, be construed as a component of general intelligence (rather in the way that the Comprehension subtest of the Wechsler scales, which tests a person’s understanding of social situations, is a component of general intelligence). Among the primary empathic abilities, three classes of ability were postulated: the social-cognitive representational abilities that are assessed by theory of mind tasks, the social-

perceptual abilities that allow us to discriminate different emotional states in other people on the basis of visual, auditory, and other sensory information, and the language-dependent social cognitive abilities that represent knowledge about emotional states, the conditions that give rise to emotional states, and allow us to use reasoning skills to make inferences about the content and causes of another person's emotional experience. The ERS were designed to assess the latter two classes of abilities --labelled emotion recognition and emotion understanding tasks, respectively. The set of ERS subtests is described next.

Emotion Recognition Tasks

Fluid Emotions Test

The Fluid Emotions Test (Dyck, Farrugia, Shochet, & Holmes-Brown, 2004) is a 32-item scale that was designed to measure the speed and accuracy with which a subject can recognise static and changing/changed facial expressions of emotion. This is a computer-presented test and items are drawn from the stimulus set developed by Matsumoto and Ekman (1995) to study the universality of facial expressions of emotion. The emotions depicted are correctly recognised by most adults, but the emotions depicted are in no case invariably recognised. Cross-cultural studies of adults have shown that there is general agreement (between 70 and 80 percent) on what emotion is being depicted by a given image (Matsumoto & Ekman). It is the disagreement between judges that makes the stimuli suitable to measure individual differences in the ability to recognise emotions: greater disagreement equates to more difficult stimuli. Similarly, the difficulty of stimuli is variable within an emotion category (fear) and across emotion categories ("contempt" is more difficult than "sadness"), adding to the potential for scaling the stimuli. Items are balanced for emotion category and ethnicity and gender of the stimulus person.

Each Fluid Emotion Test item consists of a head and shoulders picture of a Japanese or Caucasian male or female expressing one of seven emotions (anger, contempt, disgust, fear, happiness, sadness, surprise) or a neutral expression. The initial image is gradually (over four seconds) transformed by morphing software into the picture of another person expressing a different emotion. Test-takers are asked to indicate what emotion is being expressed in the initial image. After responding, the image is transformed and test-takers are asked to indicate, as quickly as they can, the second emotion being depicted. The speed of response is measured with a stop-watch. The Fluid Emotions Test yields four measures: Accuracy Scale 1 (total number correct on the pre-morph emotion), Accuracy Scale 2 (total number correct on the post-morph emotion), Speed (average response time regardless of accuracy), and Speed Given Accuracy. The Speed Given Accuracy scale is based on categorising the speed of accurate responses into one of eight categories. Response times greater than 12 seconds result in a score of 0 even though the response is accurate. Times of 9 - 12 seconds are scored 1, and each subsequent 1 second decrease in latency results in an incremental score of 1. Latencies of less than 4 seconds are scored 7.

Vocal Cues Test

The Vocal Cues Test (Dyck, Farrugia, et al., 2004) was designed to measure individual differences in the ability to recognise emotions based on tone of voice cues. The Vocal Cues Test consists of two scales that represent alternative methods for generating emotion vocalisations while excluding semantic emotion cues. The design of the Vocal Cues Test-Real Words Scale is based on the work of William and Stevens (1972) who argued that vocal stimuli should consist of a standard emotion-neutral phrase. By holding the semantic content constant, discrimination of different emotions depends on the quality of other speech characteristics. The design of the

Vocal Cues Test-Unreal Words Scale is based on the work of Dusenburg and Knowler (1939) who used gibberish and alpha-numeric stimuli in their emotion vocalisations to eliminate contamination by semantic content. In order to maintain consistency with the Fluid Emotions Test, the emotions sampled in the two Vocal Cues Test scales include anger, sadness, contempt, fear, disgust, happiness, surprise, and emotion-neutral expressions.

Stimuli were generated by actors, two women and three men, with an average of almost 11 years professional thespian experience. For each emotion, actors were asked to imagine experiencing an event that would elicit the given emotion. Events were elaborated in group discussions to identify elements of the experience that would heighten the experience of the emotion. Actors rehearsed their emotion expressions until they believed they were prepared to record their vocalisations. For the Vocal Cues Test-Real Words Scale, each actor expressed the words “I can't believe it” four times in a tone of voice appropriate to each of the eight emotion categories, resulting in 160 vocal expressions of emotion. The first recording of each emotion was subsequently discarded (practice), as was one obviously imperfect vocalisation, leaving a total item pool of 119 items. This set of items was administered, with other measures, to a small sample (n=54) of university students (Holmes-Brown, 1998). Based on the results, items that were too difficult, or unreliable, or redundant to other items were removed from the scale, leaving 45 items approximately balanced for emotion category and gender of the speaker in the final scale.

For the Vocal Cues Test-Unreal Words Scale, three forms of vocalisation—gibberish, alphabet series, and numeral series—were used. Gibberish is used as an improvisation warming-up technique in which actors speak a spontaneous invented language. Actors were free to choose which form of vocalisation they would use in

their expressions of emotion. A total of 160 vocalisations was recorded, the first recording of each emotion type was discarded, leaving 120 items. This set of items was administered, with other measures, to a small sample (n=54) of university students (Holmes-Brown, 1998). Based on the results, items that were too difficult, or unreliable, or redundant to other items were removed from the scale, leaving 43 items approximately balanced for emotion category and gender of the speaker in the final scale.

Emotion Understanding Tasks

Comprehension Test

The Comprehension Test (Dyck et al., 2001) was designed to measure the ability to predict a person's emotional response based on knowledge of the situation or context to which a person has been exposed. Items were generated to sample the range of emotions and emotion causes. Emotions included anger, fear, disgust, surprise, sadness, happiness, and contempt, social variants of the basic emotions (pride, embarrassment, shame, pity), and variations in the intensity of basic emotions (terror versus fear). Emotion causes included material causes of an emotion (loss/gain of an object), social causes of an emotion (interpersonal rejection), and intrapsychic causes of an emotion (failure to achieve one's goals). Item generation took account of cognitive (Beck, 1976), interpersonal (Kohut, 1971), and experimentally-based interactive theories of the emotions (Izard, 1993).

An initial item pool of 110 items was generated to sample emotion categories, interpersonal and individual contexts, and to balance the gender of protagonists. Each item described a situation in which a protagonist was likely to experience an emotion; the test-taker was asked to identify what emotion or emotions ("feelings") the protagonist is most likely to experience. For example, "Cathy runs across the road and

suddenly hears the screech of skidding car tires. What does Cathy feel?” Responses are scored “0” if the nominated emotion is most unlikely to result from the situation (“happy” in the example above), are scored “1” if the emotion is likely to result from the situation (“scared” in the example above), and are scored “2” if the response indicates appropriate intensity and complexity (“terrified, and then, perhaps, embarrassed,” in the example above).

The initial item set was evaluated in a series of pilot investigations. Items were read by colleagues and senior students who identified items that were ambiguous, that might be perceived as offensive, or that led to difficulties in scoring. Items that could not be adequately reworded were eliminated, resulting in 92 items being retained. These 92 items were randomly divided into two 46-item alternative forms, only one of which was used in subsequent evaluations. Administration of the 46-item test to samples of university students yielded data on item difficulty and the internal consistency of the test. Twenty-five items were selected that were (a) evenly dispersed across the range of difficulty, but with (b) some over-representation of easier items to increase the suitability of the test for young children. Based on the analysis of data from a series of preliminary investigations (see below), an 11-item short form of the test was developed that had good internal consistency and range of difficulty but was still suitable for use with young children.

Unexpected Outcomes Test

The Unexpected Outcomes Test (Dyck et al., 2001) was designed to measure a person’s ability to apply reasoning skills and knowledge of the causes of emotions to explaining apparent incongruities between an emotion-eliciting context and the emotion elicited by the context. Unexpected Outcomes Test items provide information about a situation that is likely to cause an emotional response by a protagonist (“John

finally persuades Susan to go to the movies with him.”) and indicate what emotion is experienced by the protagonist (“On the way to the movies, John can hardly contain his anger.”). In each case, the emotion is not one that would usually be expected to occur in the situation. The test-taker is asked to provide additional situational information that would make the apparent incongruity explicable. In the example above, the test-taker might suggest that “On the way to the movies, Susan has explained that she accepted the invitation because her mother has told her to be nicer to ordinary boys.” Responses are scored “2” if they provide an explicit explanation of the incongruity, and are scored “1” if the explanation is not sufficiently specific to the context of the item, requires inference by the examiner, is implausible, incomplete, or does not account for the intensity of the emotion. All other responses are scored “0”.

An initial set of 23 items was designed to reflect emotion-eliciting contexts of varying difficulty. Difficulty was principally determined by whether emotion cues were explicit (“John was angry”) or implicit (“Mary smashed her fist against the wall”) and by the interpersonal complexity of the context (from no explicit or implied interpersonal context to an implicit interpersonal context to an explicit interpersonal context). Items were evaluated in a series of informal and formal pilot studies. Initial administration of items to colleagues and senior students indicated a surfeit of too-difficult items, and also to the recognition that some items could be understood in such a way that the emotional outcome was not, in fact, unexpected. Items were reworded to reduce the likelihood of double meanings and/or made easier to reduce test difficulty. Based on analysis of data from a series of preliminary investigations (see below), a 12-item short form of the test was developed that had good internal consistency and range of difficulty but was still suitable for use with young children.

Emotion Vocabulary Test

The Emotion Vocabulary Test (Dyck et al., 2001) is a 24-item test of a person's ability to define emotion words (“What does the word ‘angry’ mean?”). Based on the recognition that emotion vocabulary represents a limit to an individual's performance on other ERS, the specific words chosen for inclusion in the Emotion Vocabulary Test are taken from the scoring keys of the other ERS. The response format of the Emotion Vocabulary Test is open-ended and, similar to the test administration procedure of standard individual intelligence tests, initial responses may be queried by the examiner in order to resolve ambiguities in the initial response. Responses are scored on a 3-point scale: a score of “0” is given for an incorrect response, a score of “1” is given for a partially correct response, and a score of “2” is given for a satisfactory response. Scoring procedures were evaluated and refined in two small-scale pilot studies of adult and adolescent samples and a 12-item short form of the test was developed based on the results of preliminary studies.

General Information

Each of the ERS requires about 10 minutes to administer or about 50 minutes in total if a set of five tests is administered. Administration time is typically briefer with younger or less able children to whom fewer items are administered because of discontinuation rules, and longer with older or more able children to whom all items are administered. Set-up times are negligible, and the only materials required are a computer screen and speakers for presenting Fluid Emotions Test and Vocal Cues Test stimuli and test forms listing emotion understanding test items and recording children's responses. Like other individually administered ability tests, administration and scoring procedures need to be rehearsed prior to use with clients or research participants.

Preliminary Investigations

Initial information on the reliability and validity of the ERS was obtained from a series of unpublished studies, including studies of children age 4 to 6 years ($n=91$; Phillips, 1997), adolescents age 14 or 15 years ($n=99$; Campbell, 1998), university students ($n=126$; McAtee, 1997) and adults ($n=66$; Ferguson, 1996). In these studies, different sub-sets and versions of ERS tests were administered as the tests were revised and new tests introduced. Results from these studies indicated that the early versions of the ERS were internally consistent (e.g., Comprehension Test: $\alpha=.66$ to $.88$; Fluid Emotions Test Accuracy 1: $\alpha=.87$ to $.91$; Unexpected Outcome Test: $\alpha=.89$ to $.94$; Emotion Vocabulary Test: $\alpha=.89$ to $.90$), measured distinct but related abilities (e.g., Pearson correlation between Comprehension Test and Fluid Emotions Test Accuracy 1: $r=.19$ to $.34$; Comprehension Test and Unexpected Outcomes Test: $r=.13$ to $.47$; Fluid Emotions Test Accuracy 1 and Unexpected Outcomes Test: $r=.09$ to $.37$), and were also related to Wechsler intelligence subtests in children (Wechsler Similarities and ERS subtests: $r=.24$ to $.52$; Wechsler Comprehension and ERS subtests: $r=.37$ to $.45$) but not adults (Wechsler Similarities and ERS subtests: $r=-.02$ to $.13$). The main value of these studies was that they provided good data on the characteristics of individual items, especially in terms of their difficulty across a broad age range, their variability, and their relationship to the total test score. This information was used to construct the final versions of the ERS, including alternate forms for some subtests. There have been no revisions to the ERS since 2001, but there have been a few revisions to the scoring keys since that time, usually as a result of test-users querying unanticipated responses.

The ERS have been used in numerous research studies but, in this article, I mainly rely on data collected in four studies investigating: the severity of deficits in social cognition among children with childhood disorders ($n=167$; including autism,

n=20, Asperger's Disorder, n=28, ADHD, n=35, intellectual disability, n=34, anxiety disorder, n=14, and no psychological disorder, n=36; Dyck et al., 2001), the severity of deficits in social cognition among children and adolescents with a sensory disorder (n=163, including deafness, n=49, blindness, n=42, and no sensory disorder, n=72; Dyck, Farrugia, et al., 2004), the relationship between achievement discrepancies and social/behavioural problems in a representative sample of children aged 3 to 14 years (n=449; Dyck, Hay, Anderson, Smith, Piek, & Hallmayer, 2004; Dyck, Piek, Hay, Smith, & Hallmayer, 2006), and the ability deficits that characterise children with developmental disorders [n=159, including autism, n=30, mixed receptive-expressive language disorder, n=30, developmental coordination disorder, n=22 (Wisdom, Dyck, Piek, Hay, & Hallmayer, 2007), intellectual disorder, n=24 (Dyck, Piek et al.), and ADHD, n=53 (Piek, Dyck, Francis, & Conwell, 2007)]. Between them, these studies provide comprehensive information on the psychometric characteristics of the ERS, most of which has not been published. Readers are referred to these earlier publications for detailed information about samples and procedures.

Inter-rater Reliability of the ERS

The inter-rater reliability of the three emotion understanding tests was assessed in three samples of children and adolescents with and without sensory disorders: 30 hearing impaired, 30 vision impaired, and 30 children with no sensory impairment were randomly selected from the total sample. These tests were selected for evaluation because the scoring procedure is more subjective than it is for the emotion recognition tests. Tests were scored independently by two raters, and Pearson correlation coefficients calculated for the two sets of ratings. The results indicate that the Comprehension Test ($r=.84$), Unexpected Outcomes Test ($r=.85$), and Emotion Vocabulary Test ($r=.94$) can be reliably scored (Dyck, Farrugia et al., 2004).

Internal Consistency of the ERS

The internal consistency of the ERS as measured by Cronbach's alpha was calculated for each test in each of the independent samples obtained in the four studies as well as the pooled samples within each study, a total of 151 coefficients. For each test, the median and mean coefficients were calculated and "outliers" were noted. In a few cases, exceptionally low coefficients were obtained and appear to indicate that the test in question is not reliable for use with the given population. For example, the alpha coefficients for the Unexpected Outcomes Test in the intellectual disability group were .22 and .14 and it may be inferred that members of this group did not have the reasoning skills necessary to complete this test. In other cases, it appears that an unusually low coefficient is an artefact. For example, an alpha coefficient for the Comprehension Test in one sample of typically developing children was .34, but was .79 in two other studies.

The results indicate that with few exceptions, the emotion understanding tests are internally consistent when used with a broad range of samples. The median (mean, number of estimates) alpha coefficient for the Emotion Vocabulary Test was .84 (.83, 18), for the Unexpected Outcomes Test it was .73 (.66, 18), and for the Comprehension Test it was .72 (.68, 18). Each of these emotion understanding tasks is least reliable when used with participants who have an intellectual disability, ADHD, or a language disorder, but with the exception of the Unexpected Outcomes Test which ought not to be administered to persons with an intellectual disability, the tests are sufficiently reliable to be used with all groups. The tests are most reliable when used with participants who are typically developing or who have autism.

For the Fluid Emotions Test subscales, the coefficients were as follows: for the Accuracy 1 scale, the median was .69 (mean=.65, number of estimates=18); for

Accuracy 2, .81 (.68, 18), for Speed, .86 (.85, 14), and for Speed Given Accuracy, .81 (.79, 17). Each of these tasks tends to be least reliable when used with participants who are deaf, who have ADHD or a motor skills disorder, or who are typically developing, and the Accuracy 1 scale and Accuracy 2 scales should only be used with deaf persons in order to calculate the Speed and Speed Given Accuracy scores. These tasks are most reliable when participants have an autism spectrum disorder or a language disorder.

For the two Vocal Cues Test subscales, the coefficients are .69 (.69, 3) for the Real Words scale and .85 (.82, 10) for the Unreal Words scale. Both of these scales are least reliable among typically developing children, among whom only the Unreal Words scale has acceptable reliability (α = .63 and .85). The Unreal Words scale is highly reliable among children with autism, an intellectual disability, language disorder, or motor skills disorder, and has acceptable reliability in samples of children who are blind or who have ADHD.

Internal Convergent Validity

As measures of different components of a hypothetical higher-order empathic ability construct, the ERS subtests were expected to share variance with each other, which would be reflected by moderate positive correlations between subtest scores and high loadings on a common latent variable in factor analyses. A conceptual distinction was drawn between emotion understanding and emotion recognition tasks, but it was not clear *a priori* whether there would be sufficient variance specific to these categories for them to be empirically distinguishable from each other by means of higher correlations between tasks within a category than across categories and by distinct latent variables in factor analyses. A question that was never explicitly asked

when the tests were designed was whether the latent structure of the ERS would be stable across development, but it was implicitly assumed that it would be stable.

Consistent with the idea that empathic abilities would continue to increase long after a child has acquired a theory of mind, all of the ERS are moderately correlated with age ($r=.56$ to $.76$) in a representative sample of children, and these age effects substantially inflate correlations among the ERS (e.g., $r=.43$ to $.71$) in samples that comprise a wide range of ages (Dyck et al., 2006). When age effects are controlled in partial correlation analyses (see Table 1), correlations between different emotion understanding tasks and between different emotion recognition tasks tend to be stronger than between emotion understanding and emotion recognition tasks. Across all ERS, the correlations are quite weak and indicate that each test is measuring a distinct ability. Indeed, the low level of shared variance across the tests prompts the question of whether these tests are all related to a latent empathic ability construct.

In considering the latent structure of the ERS, the question of the stability of this structure needs to be considered. Dyck, Piek, Kane and Patrick (2009) assessed whether there are systematic differences in relationships between intellectual, language, motor, and social cognitive abilities as a function of age, and found that across four age cohorts (3 to 5, $n=117$; 6 to 8, $n=116$; 9 to 11, $n=124$; and 12 to 14 years, $n=92$), the structure of ability in each cohort differed significantly from that of each other cohort. The trend was for increasing differentiation of ability structures as age increased but with a reversal of this trend in late childhood/early adolescence. If this pattern applies to relationships among the ERS, then the structure of the ERS needs to be assessed separately in each developmental epoch.

Principal component analyses of total scores were conducted on data from the representative sample (Dyck et al., 2006) that included seven variables, all ERS except the Speed variable, which is incorporated into the Speed Given Accuracy scale, and the Vocal Cues Test Real Words Scale, which was not administered in that study. The sample was divided into the same four age cohorts as used in the Dyck et al. (2009) study. Parallel analysis (O'Connor, 2000) was used to determine how many components to extract in each analysis. Parallel analyses indicate how many eigenvalues in each dataset exceed the eigenvalues that result from analyses of random data. Parallel analysis generates random data sets with the same dimensions (Participants X Variables) as the main analysis, conducts principal component analysis on each random data set, and specifies the mean value, and the 95th percentile value, of the 1st, 2nd, 3rd, ... eigenvalue across the random data sets. Comparison of eigenvalues from the main analysis with those at the 95th percentile in the parallel analyses (based, in this case, on 1000 principal component analyses of random data for each age group) indicates how many latent variables are unlikely to be due to chance.

Based on the results of parallel analyses, only one principal component should be extracted in the youngest group and two principal components should be extracted in the three older groups. In the 3 to 5 year old group, the first principal component accounted for 67% of total test variance and each variable had a strong loading on the component (i.e., .71 - .91). In the 6 to 8 year old group, the first principal component accounted for 38% of test variance and the second component accounted for 20% of test variance. The first component was defined by the high loadings of Fluid Emotions Test variables (.76 - .89) and the second component by the emotion understanding tests (.57 - .75) and the Vocal Cues Test (.67). In the two older groups, the first two

components represented a distinction between the emotion recognition and the emotion understanding variables. In the 9 to 11 year old group, the first component accounted for 36% of test variance and was defined by the high loadings of the four emotion recognition variables (.49 to .89) and the second component accounted for 18% of test variance and was defined by the high loadings of the emotion understanding tests (.58 to .80). Results were similar in the 12 to 14 year old group: Fluid Emotion Test variables defined the first component (.49 - .89), which accounted for 44% of test variance, and emotion understanding tasks defined the second component (.70 - .86), which accounted for 20% of test variance, and the Vocal Cues Test had weaker loadings (.22 and .33, respectively) on both components. The inconsistent loadings of the Vocal Cues Test may be due to the pseudo-linguistic character of the stimuli, which makes them resemble the more verbal emotion understanding tasks (see below). In analyses where two components were extracted, the components were weakly correlated with each other ($r=.28$, .22 and .34, respectively), suggesting the presence of a higher-order factor.

Applying the same procedures to other samples, similar results are obtained. In a mixed sample of children with a range of developmental disorders (n=162; Autistic Spectrum Disorder, n=33, 4 years, 2 months to 13 years, 3 months; Intellectual Disability, n=24, 6 years, 4 months, to 11 years, 4 months; Mixed Receptive-Expressive Language Disorder, n=30, 3 years, 10 months to 12 years, 3 months; Developmental Coordination Disorder, n=24, 5 years, 0 months to 13 years, 1 month; Attention Deficit Hyperactivity Disorder, n=53, 6 years, 11 months to 11 years, 3 months), one principal component accounts for 66% of test variance and all variables load strongly on the component (.69 - .86). In a sample of blind children (n=42, 5 years, 9 months to 16 years, 10 months), and excluding the Fluid Emotions Test, one

component accounts for 70% of test variance and loadings range from .71 to .93.

Among deaf children ($n=49$, 6 years, 9 months to 19 years, 5 months), and excluding the Vocal Cues Test, one component accounts for 62% of test variance and loadings range from .67 to .86. Finally, in a small sample of children and adolescents (mean age=12.16; $SD=3.36$) with no sensory or developmental disorder ($n=72$, 5 years, 9 months to 17 years, 10 months), two components were extracted. The first included the emotion recognition scales (loadings from .56 to .92) and accounted for 49% of test variance; the second included the emotion understanding scales (loadings from .83 to .90) and accounted for 18% of test variance. The components were moderately correlated (.44).

Overall, the results of these analyses suggest that ERS all relate to a single latent variable, namely, empathic ability. In young children and children with a sensory or developmental disorder, this latent variable is undifferentiated. From school age, the latent variable is defined by two sub-components corresponding to the distinction between emotion understanding and emotion recognition. The least stable element appears to be the ability to recognise vocal emotion cues, the task that is least reliable in typically developing children.

External Convergent Validity

Relationships with other measures of social cognition, especially theory of mind tasks, have been assessed in several samples. In typically developing children, correlations between the ERS and first and second order theory of mind tasks range from $r=.34$ to .56 in 3 to 5 year olds, but are typically not significant in older children as a result of ceiling effects related to the theory of mind tasks. Children have typically acquired a theory of mind by age 5 years and so there is little or no variability in task performance in older cohorts. With so-called advanced theory of

mind tasks that assess a child's understanding of non-literal language (e.g., irony, metaphor), there is also an age-related decrease in the strength of correlations that is comparable to those observed among the ERS. For example, correlations with the Strange Stories Test (Happé, 1994) range from $r=.37$ to $.50$ in 3 to 5 year olds, range from $.09$ to $.32$ in 6 to 8 year olds, $.17$ to $.32$ in 9 to 11 year olds, and $.04$ to $.36$ in 12 to 14 year olds. In samples of children with developmental disorders, correlations with basic ($r=.44$ to $.62$) and advanced theory of mind tests ($r=.48$ to $.76$) are moderate.

Correlations between the ERS and intelligence tests tend to be as strong as correlations among the ERS. In a representative sample of children, the magnitude of correlations between the ERS and the Wechsler Vocabulary, Information, Block Design, and Picture Completion subtests ranged from $r=.36$ to $.62$ in 3 to 5 year olds, from $r=.08$ to $.32$ in 6 to 8 year olds, $r=.09$ to $.49$ in 9 to 11 year olds, and $r=.07$ to $.52$ in 12 to 14 year olds. In general, the ERS emotion understanding tasks are more strongly related to Wechsler verbal subtests than performance ones, especially correlations between the ERS Emotion Vocabulary Test and the Wechsler Vocabulary Scale, and ERS emotion recognition tasks are more strongly related to Wechsler performance subtests than verbal ones. In one sample of children with a broad range of developmental disorders ($n=162$), with the exception of the Speed scale, correlations between the ERS and Wechsler subtests are moderate to strong, ranging from $r=.44$ to $.83$. In a second sample of children with a broad range of developmental and behavioural disorders ($n=137$), correlations between the ERS and Wechsler subtests were also moderate to strong, ranging from $r=.30$ to $.76$.

That the ERS would be related to intelligence was anticipated when these scales were conceived and constructed, but it was also anticipated that there would be

sufficient variance specific to these scales that they could not be regarded as redundant to IQ testing. To assess whether the latent variables associated with the ERS are distinguishable from IQ, additional principal component analyses of total scores were conducted that included the ERS (except Speed) and the Wechsler Vocabulary, Information, Block Design, and Picture Completion scales. In three cohorts from a representative sample (6 to 8, 9 to 11, 12 to 14), two components were extracted based on the results of parallel analyses and in each case the first component was defined by the high loadings of emotion understanding and Wechsler scales and the second component was defined by the high loadings of the emotion recognition subtests. In 3 to 5 year olds, parallel analysis indicated that only one component should be extracted, and all scales had high loadings on this component. When two components were extracted, the emotion understanding tests had substantial loadings on both the first component (with Wechsler tests) and the second component (with the emotion recognition tests). Across the analyses, the correlation between the components ranged from .32 in the oldest cohort to .65 in the youngest cohort. These results indicate that the emotion recognition subtests in particular assess an ability construct that is distinct from the verbal and performance intelligence constructs assessed by the Wechsler scales.

Sensitivity to Change

It has already been reported that the ERS are moderately to strongly related to age, which indicates that these tests are sensitive to increases in ability as a function of development (see next section). The only evidence available as to whether the ERS are sensitive to changes in ability that occur as a function of systematic training in social cognition was reported by Dyck and Denver (2003). As part of that study, brief parallel forms of the Emotion Vocabulary Test, Comprehension Test, and Fluid

Emotions Test were constructed for use as pre- and post-test measures assessing the effectiveness of deaf children's emotion recognition and emotion understanding abilities. Results of that study indicated that increases in ability were observed on the emotion understanding tests, but not the emotion recognition ones. Whether this latter result is due to poor test reliability in this sample, lack of sensitivity in the Fluid Emotions Test or inadequate instruction in the program is not known.

Australian Norms

The ERS have proved useful in research on the normal and abnormal development of children, but these scales have not been widely disseminated for use in clinical assessments of children. There may now be sufficient evidence of the reliability and validity of these scales to support their use in clinical settings. To facilitate such use, developmental norms are required. Appropriate data from a representative sample of West Australian children are available (see Table 2) but have not been previously published. Because the scores of boys and girls do not differ significantly on any test, scores have been pooled across the sexes within each age category. As Table 2 shows, with the exception of the 7 year old cohort whose mean scores appear to be anomalously low, there is a clear increase in performance across the age range for all ERS. Test materials, including scoring keys, are available from the author.

Discussion

The ERS were created to measure the ability to understand the experience of other people. When used with good theory of mind tasks, they provide comprehensive information on a child's development of social cognition. The ERS have proved especially helpful in understanding some of the deficits responsible for the social difficulties experienced by children with a range of sensory and developmental

disorders. It is not only children with an autism spectrum disorder that have marked deficits in social cognition, but also children with language or motor skills problems (Wisdom et al., 2007) or a sensory disorder (Dyck, Farrugia et al., 2004). Where deficits are observed, the ERS are also useful in determining whether the deficit is general or specific (Cummins, Piek, & Dyck, 2005) and/or whether it is proportionate to a child's general intellectual (Dyck et al., 2001) or language ability (Wisdom et al.). A competent social performance is unlikely if a child doesn't understand what makes other children happy or sad or angry or scared, or why other children behave the way they do, or what other children are experiencing.

To date, the ERS have been used mainly to test hypotheses about the presence of deficits in social cognition in various clinical groups, and it is as a result of this research that comprehensive information about the psychometric qualities of the ERS is available. This information indicates that the ERS can be reliably scored, are internally consistent, are convergent with other measures of social cognition and with general intelligence, and meaningfully distinguish emotion understanding from emotion recognition abilities. Data from a representative sample indicate how different emotion understanding abilities increase across childhood and early adolescence, and provide a good basis for evaluating whether an individual child's development of social cognition is delayed. The limited use of the ERS in treatment outcome research suggests that they are sufficiently sensitive to changes in ability to be used to assess the effectiveness of treatments designed to enhance a child's empathic abilities (Dyck & Denver, 2003).

Since the ERS were developed, other measures of emotion understanding and recognition abilities have also been developed by other researchers. Some of these tests, like the Reading the Mind in the Voice test (Golan, Baron-Cohen, Hill, &

Rutherford, 2007) and the Reading the Mind in the Eyes test (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) are quite similar to the ERS Vocal Cues and Fluid Emotions tests and are clearly as suitable as the ERS for use in research. However, what is distinctive about the ERS compared with other recently developed measures of social cognition is their comprehensiveness, the availability of norms and their demonstrated suitability for use across the full range of childhood and early adolescence. At this time, the ERS are the only measures of emotion understanding and emotion recognition abilities for which such data are available to guide the interpretation of an individual child's test results.

Acknowledgements

The Emotion Recognition Scales were planned and initially developed in collaboration with Ian Shochet and Analise O'Donovan, and several students whose work has not been cited in this paper also contributed to the first evaluations of these scales, including Stuart Bird and Julia Rudolph. Thanks to Peter Creed for his comments on an earlier version of this paper.

References

- Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, *21*, 37-46.
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (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. *Journal of Child Psychology & Psychiatry*, *42*, 241-251.
- Beck, A. (1976). *Cognitive theory and the emotional disorders*. New York: New American Library.
- Beck, A., Rush, A., Shaw, B., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.
- Campbell, I. (1998). *Empathic ability as an antecedent to prosocial behaviour*. Unpublished honours thesis, Griffith University.
- Cummins, A., Piek, J., & Dyck, M. (2005). Motor coordination, empathy and social behaviour in school aged children. *Developmental Medicine & Child Neurology*, *47*, 437-442.
- Dunn, J. (1995). Children as psychologists: The later correlates of individual differences in understanding of emotions and other minds. *Cognition & Emotion*, *9*, 187-201.
- Dusenburg, D., & Knower, F. (1939). Experimental studies of the symbolism of action and voice. II. A study of the specificity of meaning in abstract tonal symbols. *Quarterly Journal of Speech*, *25*, 67-75.
- Dyck, M., & Denver, E. (2003). Can the emotion recognition ability of deaf children be enhanced? *Journal of Deaf Studies & Deaf Education*, *8*, 348-356.

- Dyck, M., Ferguson, K., & Shochet, I. (2001). Do autism spectrum disorders differ from each other and from non-spectrum disorders on emotion recognition tests? *European Child & Adolescent Psychiatry, 10*, 105-116.
- Dyck, M., Farrugia, C., Shochet, I., & Holmes-Brown, M. (2004). How empathic ability develops in deaf or blind children: Do sounds, sights, or words make the difference? *Journal of Child Psychology & Psychiatry, 45*, 789-800.
- Dyck, M., Hay, D., Anderson, M., Smith, L., Piek, J., & Hallmayer, J. (2004). Is the discrepancy criterion for defining developmental disorders valid? *Journal of Child Psychology & Psychiatry, 45*, 979-995.
- Dyck, M., Piek, J., Hay, D., Smith, L., & Hallmayer, J. (2006). Are abilities abnormally interdependent in children with autism? *Journal of Clinical Child & Adolescent Psychology, 35*, 20-33.
- Dyck, M., Piek, J., Kane, R., & Patrick, J. (2009). How uniform is the structure of ability across childhood? *European Journal of Developmental Psychology, 6*, 432-454.
- Eisenberg, N., Murphy, B., & Shepard, S. (1996). The development of empathic accuracy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 73-115). New York: Guilford Press.
- Ferguson, K. (1996). *Development and preliminary validation of the "Unexpected Outcomes Test"*. Unpublished graduate diploma dissertation, Griffith University.
- Flavell, J. (1992). Perspectives on perspective taking. In H. Beilin & P. Pufall (Eds.), *Piaget's theory: Prospects and possibilities* (pp. 107-139). Hillsdale, NJ: Erlbaum.

- Gillberg, C. (1992). Autism and autistic-like conditions: Subclasses among disorders of empathy. *Journal of Child Psychology & Psychiatry*, 33, 813-842.
- Gillberg, C. (1993). Autism and related behaviours. *Journal of Intellectual Disability Research*, 37, 343-372.
- Golan, O., Baron-Cohen, S., Hill, J., & Rutherford, M. (2007). The “Reading the Mind in the Voice” Test - Revised: A study of complex emotion recognition in adults with and without autism spectrum conditions. *Journal of Autism & Developmental Disorders*, 37, 1096-1106.
- Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters’ thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism & Developmental Disorders*, 24, 129-154.
- Holmes-Brown, M. (1998). *The development and preliminary validation of the Vocal Cues Test: Measuring emotion recognition*. Unpublished honours thesis, Griffith University.
- Izard, C. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, 100, 68-90.
- Kohut, H. (1971). *The analysis of self*. New York: International Universities Press.
- Matsumoto, D., & Ekman, P. (1995). *Japanese And Caucasian Facial Expressions Of Emotion (JACFEE) And Neutral Faces (JACNeuF)*. San Francisco: San Francisco State University.
- McAtee, E. (1997). *The empathic ability of repressors vs nonrepressors*. Unpublished honours thesis, Griffith University.

- O'Connor, B. P. (2000). SPSS, SAS, and MATLAB Programs for Determining the Number of Components Using Parallel Analysis and Velicer's MAP Test. *Behavior Research Methods, Instruments & Computers*, 32, 396-402.
- Phillips, L. (1997). *What is empathy? The demystification of empathy and the development of a new empathy measure*. Unpublished masters dissertation, Griffith University.
- Piek, J., Dyck, M., Francis, M., & Conwell, A. (2007). Working memory, processing speed and set-shifting in children with Developmental Coordination Disorder and Attention Deficit Hyperactivity Disorder. *Developmental Medicine & Child Neurology*, 49, 678-683.
- Rogers, C. R. (1951). *Client-centered therapy*. Boston: Houghton-Mifflin.
- William, C., & Stevens, K. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52, 345-350.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition*, 13, 103-128.
- Wisdom, S., Dyck, M., Piek, J., Hay, D., & Hallmayer, J. (2007). Can autism, language and coordination disorders be differentiated based on characteristic ability profiles? *European Child & Adolescent Psychiatry*, 16, 178-186.

Table 1

Partial correlations among ERS controlling for age

	2	3	4	5	6	7	8
1 Comprehension Test	.22***	.34***	.25***	.24***	-.09	.20***	.30***
2 Unexpected Outcomes Test		.36***	.12**	.12*	-.15**	.21***	.23***
3 Emotion Vocabulary Test			.09	.14**	-.12**	.16***	.22***
4 Accuracy 1				.81***	-.21***	.59***	.39***
5 Accuracy 2					-.25***	.72***	.41***
6 Speed						-.61***	-.21***
7 Speed Given Accuracy							.30***
8 Vocal Cues Test Unreal Words							

 $df=418$

* Coefficient is significant at the .05 level, two-tailed

** Coefficient is significant at the .01 level, two-tailed

*** Coefficient is significant at the .001 level, two-tailed

Table 2

Means (and standard deviations) of ERS by age in years

Age (n)*	CT	UOT	EVT	ACC1	ACC2	SPD	SGA	VCTU
3 (20)	3.19 (1.88)	1.64 (1.72)	2.35 (1.79)	7.40 (3.92)	7.60 (4.10)	8.73 (2.93)	24.35 (16.48)	8.10 (7.29)
4 (37)	4.40 (2.25)	2.30 (1.77)	4.02 (2.88)	9.62 (5.53)	10.10 (6.18)	8.44 (2.01)	33.37 (23.63)	10.07 (7.93)
5 (41)	6.75 (2.23)	4.50 (1.88)	7.37 (2.90)	14.45 (4.23)	14.97 (4.32)	6.91 (0.74)	54.80 (19.33)	19.54 (6.28)
6 (34)	8.20 (2.12)	5.20 (2.56)	9.51 (3.83)	18.58 (2.76)	17.55 (2.83)	6.46 (0.97)	70.38 (19.71)	19.54 (5.06)
7 (40)	7.43 (2.41)	3.52 (2.33)	5.46 (2.84)	17.70 (3.42)	17.70 (2.79)	6.82 (1.60)	72.30 (16.78)	17.50 (4.43)
8 (40)	9.05 (1.85)	4.95 (2.97)	7.65 (2.80)	19.62 (2.59)	19.00 (3.31)	6.42 (1.41)	82.50 (20.90)	19.60 (4.53)
9 (40)	9.53 (2.49)	7.87 (3.75)	8.90 (3.43)	19.87 (2.90)	19.31 (3.25)	6.30 (1.35)	85.24 (19.61)	20.05 (4.03)
10 (42)	9.68 (2.33)	6.06 (3.43)	10.82 (3.26)	19.02 (2.98)	19.23 (3.18)	5.75 (1.02)	90.20 (18.59)	21.23 (3.73)
11 (38)	10.61 (2.34)	7.89 (3.64)	12.31 (3.35)	20.84 (2.93)	21.05 (3.12)	5.77 (0.86)	96.31 (18.12)	22.55 (4.11)
12 (35)	11.39 (2.43)	9.54 (4.23)	15.16 (4.43)	19.67 (4.16)	20.24 (3.75)	5.57 (0.91)	98.32 (24.42)	22.94 (5.33)
13 (35)	12.80 (2.54)	8.20 (3.58)	15.26 (4.08)	22.11 (3.90)	21.77 (3.33)	5.34 (1.07)	110.51 (25.98)	24.45 (4.48)

14 (19)	13.78 (2.50)	12.42 (4.92)	18.42 (4.46)	22.15 (2.71)	22.42 (3.61)	4.75 (0.79)	123.26 (26.48)	26.63 (4.13)
---------	--------------	--------------	--------------	--------------	--------------	-------------	----------------	--------------

* In some age cohorts, the number of cases differs across variables; the n provided is the minimum number of cases available.

Abbreviations: CT=Comprehension Test; UOT=Unexpected Outcomes Test; EVT=Emotion Vocabulary Test; ACC1=Fluid Emotions Test

Accuracy 1 Scale; ACC2=Fluid Emotions Test Accuracy 2 Scale; SPD=Fluid Emotions Test Speed Scale; SGA=Fluid Emotions Test Speed

Given Accuracy Scale; VCTU=Vocal Cues Test Unreal Words Scale.