Internet Interventions 24 (2021) 100381

Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.elsevier.com/locate/invent

Co-design and qualitative validation of animated assessment item content for a child-reported digital distress screener

Kirsty Zieschank [a,*], Jamin Day [b], Michael J. Ireland [a], Sonja March [a]

[a] School of Psychology and Counselling and Centre for Health Research, University of Southern Queensland, Springfield Campus, PO Box 4196, Springfield Central, Queensland 4300, Australia
[b] Family Action Centre, The University of Newcastle, University Drive, Callaghan, NSW, Australia

ARTICLE INFO

Keywords:
Child codesign
Digital screener
Iterative development
Animation

ABSTRACT

Purpose: The *Interactive Child Distress Screener* (*ICDS*) is a novel, digital screening tool that is currently under development and aims to broadly assess self-reported emotional and behavioural distress in children aged five to 11 years. This study implemented a generative participatory codesign and iterative refinement process to qualitatively validate the content of 30 animated assessment items developed for the ICDS by assessing their acceptability and accuracy from the child's perspective.
Methods: The participants ($N = 62$) were five to 11-year-old children. Individual interviews were conducted with each child to determine acceptability and validity of animated items and facilitate the co-design refinement process of the animated assessment items.
Results: Twenty-two out of 30 (73%) items met ≥80% satisfaction and accuracy consensus in their original format, six items (20%) required one round of refinement before meeting consensus, and two items (7%) required two rounds of refinements. Combined acceptability of animated items was high, ranging from 4.1 to 5 out of 5 across all items.
Conclusion: Participants were able to accurately identify and understand socio-emotional and behavioural constructs when depicted as animated items. Acceptability was high, even in first iterations when accuracy of understanding required refinement. This study highlighted the importance and benefits of iterative participatory design methodology in ensuring assessment items developed for children are understood, accepted and likely to be effective in obtaining accurate self-report.

1. Introduction

1.1. Background

According to national mental health surveys (Lawrence et al., 2016; Sawyer et al., 2001) emotional and behavioural disorders remain the most commonly diagnosed among children under the age of 12 years and include attention deficit/hyperactivity disorder (ADHD, 7.4%), anxiety disorder (6.9%), major depressive disorder (MDD, 2.8%), and conduct disorder (2.1%). Comorbid disorders are frequent, with one third of the children diagnosed with ADHD or conduct disorder also suffering from anxiety and/or depression (Achenbach and Rescorla, 2003; Johnson et al., 2016; Wood and McDaniel, 2020). Undetected and untreated mental illness during childhood causes suffering, impedes healthy development, has detrimental effects on educational progress and opportunities through to adolescence, and also increases the probability of enduring psychosocial disorders in adulthood (Caspi et al., 1995; Moreira et al., 2013; The Royal Australian and New Zealand College of Psychiatrists, 2010). The early detection of symptoms of distress is key, as many emotional and behavioural difficulties can be treated effectively if identified early and before they develop in intensity (Jacka and Reavley, 2014).

Universal and targeted mental health screening as a first step in early intervention for children has long been a priority recommended by Government and professional mental health bodies, but in practice, is not broadly implemented (Children's Health Queensland, 2018; Royal Australian and New Zealand College of Psychiatrists (RANZCP), 2017). Pre-emptive screening provides the means to identify those children showing early symptoms: initially for the purposes of referral for comprehensive assessment which facilitates early intervention, and

* Corresponding author.
E-mail address: kirsty.zieschank@usq.edu.au (K. Zieschank).

https://doi.org/10.1016/j.invent.2021.100381
Received 19 June 2020; Received in revised form 12 February 2021; Accepted 22 February 2021
Available online 4 March 2021

ultimately to reduce substantial mental health and financial cost burdens before they begin. However, there are distinct challenges to this. To have full impact, mental health screening must be equitably accessible and gather perspectives from multiple responders including the child.

The conventional approach to examining children's experiences has been to observe them and make subjective judgements or rely primarily on proxies such as parents, clinicians, or teachers to respond on their behalf (Darbyshire et al., 2005). This is despite support in the literature establishing the reliability and validity of children's ability to self-report, particularly as related to their internal subjective experiences which are more easily hidden (Cree et al., 2002; Hudziak et al., 2007; Kirk, 2007; Riley, 2004). Parent reports may be influenced by their predominant concerns, their level of involvement in primary caregiving, and their own mental health and wellbeing, whilst clinician and teacher perspectives are restricted to the settings in which they see the child (Eiser and Morse, 2001). To ensure that a comprehensive picture is obtained, it is important to have screening tools that offer children a self-report option. Yet, screeners that have sufficient evidence base to warrant widespread use are costly and/or require professional administration which limits their accessibility or offer self-report options for adolescents only which overlooks younger children's perspectives. For example, the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) and the Brief Problem Monitor-Youth Form (BPM-Y; Achenbach et al., 2011) are two well validated screening tools that are restricted to clinical settings and do not have a self-report option for children <11 years.

Further, there is growing evidence demonstrating discrepancies between parent-child agreement ratings across multiple measures and particularly within emotional and psychosocial domains (Jardine et al., 2014). Riley (2004) reported a meta-analysis of 119 studies that found the average correlation between parent and teacher reports and parent and child reports (when the option was available) were 0.28 and 0.22 respectively. A systematic review examining agreement between self- and proxy-reported quality of life (e.g., anxiety, pain, depression, coping) in young children aged <12 years found that the child's perception frequently differed from their parents in both positive and negative directions (Jardine et al., 2014). Such discrepancies might be associated with the mode of delivery or the way in which measures have been written for children. To meet such challenges, new methods for developing and delivering screening instruments for children are needed to increase accessibility and to supplement proxy reports so that a more comprehensive understanding of a child's mental health status might be consistently obtained.

1.2. Instrument development for children

Despite assessment being an integral component of clinical psychology practice, there is no 'gold standard' approach for developing measurement instruments for children (Bergeron et al., 2013). Screening instruments developed for emotional and behavioural assessment are typically modified adult, pen-and-paper, text-based measures and contain words that may be problematic for younger children to understand (e.g., depressed, inferior, self-conscious, stubborn). Further, they often utilize Likert-style response formats with three or more response options. Mellor and Moore (2014) found that dichotomous response formats were more reliable than Likert scales with children aged 6–13 years, especially when used with questions concerning emotional or behavioural states.

Understanding how children internalize and comprehend assessment item content and formulate responses is an imperative first step in designing sound instruments for children for two key reasons. First, this maximizes item response accuracy; and second, it increases clinical interpretability of test results and informs how these results are used to make clinical recommendations. Therefore, self-report measures for children must incorporate new methodological approaches that emphasize the active participation of young people (i.e., co-design methods) to accurately capture the child's perspective.

1.3. Obtaining the child's perspective via participatory methods

The practical-methodological justification for including children in child-focused research is because they know most about their lives and are therefore, the best sources of information (Soffer and Ben-Arieh, 2014). Generative, participatory research design means children are involved in all creative development activities that facilitates their contribution as research partners, and not simply as observed subjects or testers of a final product (Stålberg et al., 2016; Vandekerckhove et al., 2020). In relation to research involving digital technology, understanding the child's perspective benefits from participatory co-design approaches that are characterized by iterative phases (Mummah et al., 2016; Stålberg et al., 2016; Stoyanov et al., 2016). A true iterative process requires cyclical inquiry where evaluations, revisions, and improvements are the outcome of each iteration until a conclusion is reached (Cockburn, 2008; Farcic, 2014; Patrick et al., 2016). In iterative methodologies involving children, the child's perspective is valued throughout the development process because their input is essential at each improvement phase, and ultimately increases the likelihood of their engagement with the end-product (Edwards et al., 2016; Stålberg et al., 2016).

Though this approach is still quite unique with respect to the development of child mental health assessment instruments, the use of participatory and iterative development techniques has been employed with children in positive psychology research and with adolescents when composing questionnaires, (Ten Brummelaar et al., 2014; Yarosh and Schueller, 2017). In the case of developing digital tools, this type of methodology involves drawing on a range of interviewing techniques that include prompting spontaneous narratives, soliciting responses to scenarios and vignettes, using visual and audible stimuli, and obtaining consultative feedback (Greene and Hogan, 2005).

1.4. The digital opportunity

Low-cost internet has made digital health and psychological treatment interventions via website and contemporary mobile application software increasingly accessible over the past ten years (Cugelman, 2013; Marsac et al., 2015; Newton et al., 2016). It follows that digital technology will also increase mental health screening opportunities and overcome many inherent challenges of widespread screening of young populations. Attempts have previously been made to improve on traditional paper-and-pencil assessments for children via the addition of digital images alongside written questions. For example, desktop computer-based versions of the Dominic Interactive (DI), Strengths and Difficulties Questionnaire (SDQ), and Mood Assessment via Animated Characters (MAAC), were all trialed in the decade preceding 2010 to increase user engagement and understanding (Manassis et al., 2009; Truman et al., 2003; Valla et al., 2000). Limited psychometric data and information on the development of these measures is available and none appear to be in use today. Initial studies variously reported high user satisfaction and some clinical utility (i.e., computerized SDQ), moderate convergent validity (i.e., DI), and discrimination between anxious and nonanxious children (i.e., MAAC). Notably, all demonstrated improved engagement compared to standard pencil-and-paper versions suggesting visual, digital formats are favored by children (Bergeron et al., 2013; Linares Scott et al., 2006; Manassis et al., 2013; Truman et al., 2003).

The current generation of children are exposed to digital technologies from a very young age and learn to independently access and operate websites and programs by simply tapping and swiping on touchscreen devices with a finger (Wrobel, 2019). Consequently, we propose that digital instruments comprised of child-friendly assessment items and response modalities that children are already accustomed to will obtain reliable clinical information and provide means for accessible and rapid screening. Modern animation techniques offer a novel

and promising approach to improve on the static images used in previous efforts. Audio-visual assessment items might better facilitate accurate child-report, particularly in contexts where standard written question and answer assessments are unsuitable or self-report options are currently non-existent for children under 12 years of age. To ensure that animated item content is meaningful to children and accurately demonstrates socio-emotional and behavioural concepts from the child perspective, we further propose that children must be involved in item development using generative participatory research methodologies.

1.5. The Interactive Child Distress Screener (ICDS)

It is intended that the ICDS will be delivered via a user-friendly web-based app and utilized as a broad screening instrument designed to detect self-reported emotional and behavioural difficulties in children aged 5 to 11 years. A prototype version that includes three pairs of original animated cartoon assessment items has been tested by children, with results highlighting the acceptability and feasibility of the animated format (March et al., 2018). Ultimately, the ICDS will utilize up to 15 pairs of contrasting digital animations as assessment items: each representing different socio-emotional or behavioural situations. The broad social-emotional and behavioural domains and 15 associated constructs were selected by an expert panel in an original pilot study (March et al., 2018).

A follow-on qualitative focus-group study was conducted with 20 children to explore how they understood, visualised, and expressed each of the 15 emotional and behavioural constructs that the proposed ICDS would measure (Zieschank et al., 2020). Participatory methodologies were implemented to engage children in role play and discussion to provide visual, verbal, and physical interpretations of each construct item to capture the child's perspective. Comprehensive typologies of each emotion and behavior were created from this data to guide the translation of the children's collective viewpoint into 15 pairs of animated prototype screening items. Results of this study also demonstrated the importance of audio-visual depiction over simple lexical labelling and a lack of distinct developmental differences in emotion comprehension and expression between younger and older participants (Zieschank et al., 2020).

1.6. Aims

This paper describes the generative participatory methodology and iterative process that was implemented to develop 30 prototype cartoon animations (i.e., 15 contrasting pairs) to be used as assessment items in the ICDS instrument. The broad objective of the present study was to co-design and refine the animated items with children (5–11 years) via qualitative interviews. Specific aims were to 1) determine the acceptability of the prototype animations to children via satisfaction ratings and 2) to conduct qualitative validation of animation content by assessing the accuracy of children's understanding and recognition of the target and contrasting social, emotional, and behavioural constructs. Participatory and iterative methodologies were utilized to conduct this research until optimal acceptability and accuracy of animated item content was reached. This study received approval from The University of Southern Queensland's Human Research Ethics Committee (Ref: H16REA003).

2. Materials and methods

2.1. Materials

2.1.1. 30 prototype digital animations (ICDS items)

A set of 30 audio-visual animations displayed in MPEG-4 AVC video file format (Mp4) was created using 2D Vector-based animation techniques. The findings from previous focus groups with children (Zieschank et al., 2020) were used to inform the original scenario content for each of the items. The contrasting animation pairs that form each assessment item are listed in Table 1. Target animations depict a child experiencing a difficult emotional or behavioural experience (e.g., sadness) whilst the contrasting item depicts the opposite of each target construct (e.g., happiness).

2.1.2. Interview script

The interview script was developed specifically for this study and comprised a series of questions to qualitatively examine the acceptability and content accuracy of the animated items and invite refinement suggestions for improvement.

2.1.2.1. Animation acceptability questions. Participants' acceptability of each animation was determined via satisfaction ratings obtained using a 4-item author-developed survey. For each animation, participants were asked to rate how much they liked 1) the sounds heard during each animation (audio appeal), 2) the animated characters (character appeal), 3) the actual animated scene (animated action and context), and 4) the animation overall (viewing appeal). Responses were elicited using a 5-point visual Likert scale which utilized stars rather than numbers and labels. A greater number of stars equated to higher satisfaction, such that one star indicated the lowest satisfaction (scored as one point), and five stars indicated the highest satisfaction (scored as five points). Scores for each animation were averaged across the four categories to provide a total acceptability rating out of five. The higher the score, the higher the satisfaction and therefore the acceptability of the animation to the participant. Average satisfaction scores (out of five) were calculated for each of the acceptability questions.

2.1.2.2. Animation accuracy questions. The intended emotion or behavior being conveyed in each animation was not divulged to the participant until they had answered the first two accuracy questions. The aim of accuracy questions was to ascertain the participants' ability to correctly understand the intended construct depicted in the animation in four ways. To determine if each participant could: 1) accurately understand the intention of the animation, 2) accurately identify the construct depicted in the animation by verbally labelling it, 3) judge whether they believed the animation content accurately represented the construct as intended, and 4) judge whether the audio soundtrack accurately enhanced the animation content. Based on aggregated participant responses, the animation content was refined at each round.

Table 1
Target and contrasting constructs for which 15 animated assessment item pairs were created.

Item	Animation	Target construct	Animation	Contrasting construct
1	1a	Sad – Depressed	1b	Happy
2	2a	Worried – Anxious	2b	Not worried – Confident
3	3a	Sleeps poorly	3b	Sleeps well
4	4a	Angry	4b	Not angry – Impassive
5	5a	Disobedient (at School)	5b	Obedient (at School)
6	6a	Shy	6b	Not Shy – Outgoing
7	7a	Argumentative	7b	Not argumentative
8	8a	Hyperactive behavior	8b	Calm – Sensible
9	9a	Lonely – Alone	9b	Sociable – Alone by choice
10	10a	Bullied – Excluded	10b	Not bullied – Included
11	11a	Fearful – Scared	11b	Not Scared – Brave
12	12a	Disobedient (at Home)	12b	Obedient (at Home)
13	13a	Distracted – Inattentive	13b	Focused – Pays attention
14	14a	Physically aggressive	14b	Kind – Peaceful
15	15a	Physical symptoms – Feel sickly	15b	No physical symptoms – Feel well

2.1.2.2.1. Question 1. Understanding. To determine how accurately the children understood the content of each animation they were asked to recall a personal account equivalent to the animation content with the question "Tell me a story about a time when you or someone you know felt the same as the child in this cartoon?". To be rated as correct, the investigator considered whether a participant's narrative example was comparable to the emotion or behavior depicted in the animation. For example, a narrative about a beloved pet dying and a parent being hospitalized were recognized as scenarios that would elicit sadness and worry. These narrative examples were deemed comparable, internalized understandings of animations 1a (sad-depressed) and 2a (worried-anxious) and scored as correct. All responses were recorded verbatim on the response sheet and coded as either correct (1) or incorrect (0). A correct response indicated that the child understood the animation content and that the animation was accurate in its depiction, and an incorrect response meant they misunderstood the intention of the animation which might require refinement. Two assessors independently reviewed participants narratives to evaluate if the child's internal representation (understanding) accurately aligned with the intended construct. Any discrepancies were discussed until consensus was reached.

2.1.2.2.2. Question 2. Identification. To determine how accurately children identified the construct depicted in each animation they were asked to verbally label the emotion or behavior with the question "Can you tell me how the child is feeling or behaving in this cartoon?". Participants' labelling ability was verbally affirmed if accurate or corrected if their interpretation was inaccurate. For example, responses such as 'sad', 'upset', or 'unhappy' were scored as accurately identifying item 1a (sad-depressed). Responses were recorded verbatim and coded as correct (1) or incorrect (0). Two assessors independently reviewed all participant responses to evaluate if the child's lexical descriptor accurately identified each animation as intended. Any inconsistency between investigator ratings was discussed between assessors until consensus was reached.

2.1.2.2.3. Question 3. Representation. To determine if participants believed the animation in question was an accurate representation of the intended construct, they were asked the question "Is this animation good or bad at showing someone feeling (e.g., sad)?". To answer this question participants were provided with a sheet of paper displaying a large red cross symbol and a large green tick symbol. Children responded by pointing to their chosen symbol or by saying yes, no, tick, or cross. This question was asked after the child was informed of the construct that the animation was meant to portray. If a participant responded in the affirmative and deemed the animation to be 'good', it was coded (1) as an accurate representation. If they deemed the animation to be 'bad' and responded negatively, the animation was considered to be inaccurate representation of the intended construct and was coded (0).

2.1.2.2.4. Question 4. Audio soundtrack. To determine if the audio soundtrack accurately enhanced their understanding of the emotion, behavior, or scenario depicted in each animation, participants were asked "Did the sounds help you to understand how the child was feeling and what was happening in the cartoon?" A "yes" response was coded (1) for helpful sounds that aided understanding of the animation, and "no" was coded (0) for sounds that were not helpful or confused their understanding of the animation.

2.1.2.3. Item refinement questions. To aid refinement of items, participants were asked to first offer a 'better idea' or 'different story' for any animation they believed inaccurately represented a construct by responding to the question "What could we do in this cartoon so that it does show 'x' feeling or behavior?". Participants were asked to provide qualitative feedback about the scenario and discuss suggestions on how to change the animation to increase the accuracy of the content. They were asked about adding, removing, enhancing, or changing the story, action, background scene, or sounds. Responses were recorded verbatim onto an answer sheet and assigned according to the participants', 'add', 'remove', or 'change' recommendations. Any item identified as requiring refinement was redeveloped based on this feedback in conjunction with discussion with the animator.

2.2. Sample

Data were collected from a community sample of 62 children (50% male, $M_{age} = 8.10$ years, $SD = 2.00$) living in South-East Queensland, Australia. Participants were contacted via their parents through the research team's personal networks and via advertisements on social media. Inclusion criteria was that children were aged from five to 11 years, spoke English, and were able to attend an in-person interview. No inclusion or exclusion criteria around ethnicity, mental health difficulties or emotional or behavioural symptoms were included or assessed. Gender ratios were roughly equivalent across ages, as outlined in Table 2. All children were born in Australia.

2.3. Procedure

Parents who responded favorably to advertisements were emailed comprehensive information sheets describing the purpose of the study, their child's right to decline or withdraw their consent to participate at any time and the confidentiality of their child's responses. With parents' written consent, the first author met with each child individually, reiterated the content of the information sheet, and obtained their written assent to participate. Structured, individual interviews took place at the University of Southern Queensland or at the child's own home at the preference of the participant's parent. The interviews were conducted using a semi-structured script. Participants were asked to watch a single animation and then answer the interview questions in order. This process was repeated until each animation had been viewed and examined by the participant and responses recorded. Acceptability and accuracy questions were asked verbatim to ensure standardization across participant interviews, whilst animation refinement discussions were less structured. Participants were able to take breaks and could view any animation multiple times, therefore the length of interviews varied between participants. All children were provided with a store gift card to the value of $20 AU at the completion of each interview to thank them for their time. Children had no prior knowledge that a reward would be provided.

To ensure young participants were not unduly burdened by overly lengthy interviews, the maximum number of animations that an individual participant examined was eight pairs (16 discrete animations). This was achieved by splitting participants into pools within each round of interviews. Every attempt was made to ensure similar age and gender representations within each participant pool. However, given that it was a convenience sample, this was dependent on the availability of participants at the time the interviews were conducted. Fig. 1 describes the group characteristics of each interview pool and the iterative development phase they were involved in. Interviews were conducted over a six-month timeframe to allow for refinement of animations by the animator between interview rounds. All refinements to animations were completed by a digital animation artist based on updated storyboards and their ability to be accurately animated. Refinements were grounded in aggregated participant feedback suggestions. At least 80% of participants perceived evaluated animations to be accurate in their original

Table 2
Total number of participants as a function of age and gender (N = 62).

	Child age (years)							
	5	6	7	8	9	10	11	Total (n)
Gender								
Male	4	3	5	4	3	3	3	31
Female	5	4	4	6	4	9	5	31

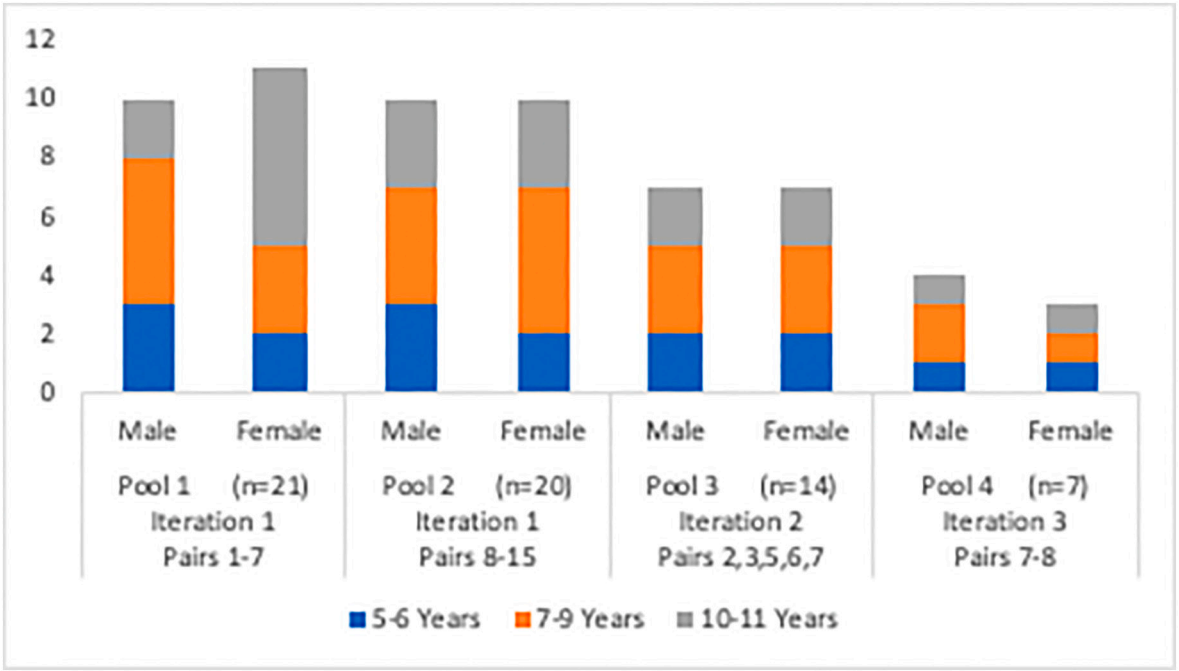

Fig. 1. Age and gender characteristics of participants per interview pool and iteration round (N = 62).

prototype format, therefore fewer participants were required for following iterations.

2.4. Analytic approach

To determine if an animation required refinement, acceptability ratings, accuracy consensus ratings, and participant refinement suggestions were examined per participant pool. Satisfaction ratings and interview responses (coded as correct or incorrect) were analyzed after each interview round to determine which animations were accurately interpreted and acceptable to participants across all ages (5–11-years) and within age group levels (i.e., level one = 5–6-years, level two = 7–9-years, and level three = 10–11-years). A consensus method was applied (determined a priori) such that for an animation to be retained, it needed to: a) receive at least four stars on average (out of five) on total average satisfaction scores to be deemed acceptable, and b) achieve at least an 80% rating consensus across the whole sample (5–11-years) within 'understanding', 'identification', 'representation', and 'audio' accuracy assessment response categories. Any animations that failed to reach consensus across the whole sample in any response category were identified for refinement and further examination in a following iteration. Animations that reached consensus across the sample as a whole but failed within age-group levels in one or more categories were considered for refinement based on individual responses and participant refinement suggestions. Further, if participants gave substantial suggestions for improvements to an animation despite high satisfaction and accuracy ratings, then the animation was likewise considered for refinement to increase acceptability.

Refined animations were then presented to a new participant pool in a subsequent interview round for re-evaluation. Refined animations were always presented with their paired contrasting animation (even if this animation had not been refined) to ensure context was preserved and so that all participants viewed animations consistently across participant pools.

3. Results

Results are reported sequentially by interview iteration and participant pool. The specific protype animated items being examined are reported, followed by acceptability rating results, accuracy consensus ratings, and refinement suggestions. A results summery details which animated items were retained, and which were marked for refinement and examination in a following interview round.

3.1. First iteration results

The 15 original pairs of animations were divided among two interview pools in round one interviews. Pool one participants (n = 21, 52.4% female, M_{age} = 8.14 years, SD = 1.96) examined 14 individual animations (i.e., item pairs 1 to 7) and pool two participants (n = 20, 50% female, M_{age} = 8.10 years, SD = 2.10) examined 16 animations (i.e., item pairs 8 to 15).

3.1.1. Animation acceptability

The acceptability of all animations was high with 100% rated ≥4.0 stars. The average satisfaction ratings for each item out of five stars are reported in Table 3. In participant pool one, the average star-rating across all animations for each response category was 4.1 for audio appeal, 4.3 for character appeal, 4.2 for animated action and context, and 4.4 for viewing appeal. Combined ratings were high across all animations with an average total satisfaction score of 4.3 representing positive acceptability. From participant pool two, the average star-rating across all animations for each category was 3.7 for audio appeal, 4.3 for character appeal, 4.2 for animated action and context, and 4.5 for viewing appeal. The average rating for audio appeal was lower across all animations in this pool due to one participant consistently rating the audio appeal of most items as one star. Despite this, the combined satisfaction ratings were high across all animations with an average total satisfaction score of ≥4.1.

3.1.2. Animation accuracy

Five items failed to reach consensus (i.e., ≥80% accuracy) across the whole sample in one or more response categories and therefore immediately identified for refinement. These were items 5b, 6a, 7a, 7b, and 8a. The poorest performing item was 7b (not argumentative) with 43% of participants misunderstanding the intended behavior and responding

Table 3
Iteration one: average satisfaction ratings (out of 5) per item.

Animated Item	Audio Appeal	Character Appeal	Animated Action Context	Viewing Appeal	Total Average Satisfaction
Pool 1 (N = 21)					
1a	4.0	4.3	4.0	4.4	4.2
1b	3.9	4.0	4.4	4.5	4.2
2a	4.1	4.1	4.2	4.2	4.2
2b	3.9	4.1	4.1	4.4	4.1
3a	4.2	4.3	4.4	4.6	4.4
3b	4.2	4.4	4.1	4.4	4.3
4a	4.4	4.3	4.4	4.4	4.4
4b	4.0	4.1	4.1	4.1	4.1
5a	4.3	4.2	4.3	4.6	4.4
5b	4.1	4.3	4.2	4.4	4.3
6a	4.0	4.2	4.4	4.4	4.3
6b	4.0	4.7	4.2	4.6	4.4
7a	4.2	4.2	4.2	4.7	4.3
7b	3.8	4.3	4.1	4.5	4.2
Pool 2 (N = 20)					
8a	3.6	4.1	4.1	4.4	4.0
8b	3.5	4.2	4.1	4.4	4.0
9a	3.7	4.3	4.1	4.4	4.1
9b	3.6	4.2	4.1	4.4	4.1
10a	3.5	4.2	4.0	4.4	4.0
10b	3.5	4.1	4.0	4.4	4.0
11a	3.5	4.2	4.1	4.4	4.0
11b	3.5	4.2	4.1	4.4	4.0
12a	3.6	4.3	4.2	4.6	4.2
12b	3.7	4.3	4.2	4.5	4.2
13a	3.8	4.3	4.2	4.6	4.2
13b	3.8	4.4	4.3	4.5	4.2
14a	3.9	4.3	4.3	4.6	4.3
14b	3.9	4.3	4.2	4.5	4.3
15a	3.9	4.3	4.2	4.5	4.2
15b	3.9	4.3	4.3	4.6	4.3

with inaccurate personal examples, and 66% incorrectly identifying the item, with the majority labelling the scenario as "being a good girl" and representative of a child demonstrating 'good' behavior.

Though all other items achieved ≥80% accuracy consensus across the sample as a whole in each response category, there were variations within age-group levels. Items 4b, 6b, 8b, 10a, 11b, 12a, and 13a failed to reach consensus within the 'understanding' response category in one or more age-group levels (i.e., item 4b, age-group levels 1 and 2; item 6b, level 3; items 8b and 13a, level 1; and items 10a, 11b, and 12a, level 2). Closer examination of individual participant responses regarding these animations showed that discrepancies were predominantly due to children not being able to think of an equivalent personal example and replying with statements such as "I don't know" and "I don't act like that" or shrugging their shoulders, rather than incorrect responses. Consensus was achieved in all other response categories for these animations with 96.5% of participants providing accurate labels (i.e., 'identification' category), 98.6% of participants agreeing the animations accurately portrayed the constructs ('representation' category), and 100% of participants agreeing the sounds were accurate ('audio' category). No suggestions were made for refinements; therefore, a decision was made to retain these seven animations in their original form.

Item 2b (not worried – confident) achieved ≥87.5% consensus in the 'audio' category but failed to achieve consensus within age group level 2 (75%) in the 'understanding' and 'representation' categories, and age group level 3 (75%) in the 'identification' category along with numerous refinement suggestions. Item 5a (disobedient at school) achieved 100% consensus in the 'audio' category, and ≥ 80% consensus in 'understanding' and identification categories but failed to achieve consensus within age group level 3 (50%) in the 'representation' category along with several refinement suggestions. Therefore, items 2b and 5a were also targeted for refinement. Table 4 identifies accuracy ratings across the whole sample (5–11-years) for each accuracy response category (i.e., understanding, identification, representation, and audio accuracy) during Iteration One. Table 5 specifies accuracy rating variations between age group levels (level 1 = 5–6 years, level 2 = 7–9 years, and level 3 = 10–11 years).

3.1.3. Refinement feedback for identified animations

The majority of suggestions to improve items were to exaggerate components of the current animation or add in creative details. If a suggestion could not be animated effectively it was not incorporated. Any suggestions to add substantial dialogue between characters were disregarded due to the brevity of the animations and due to our prior commitment to develop measurement items that were not contingent on language (written or spoken). Though item 3a (sleeps poorly) achieved >90% consensus in each accuracy response category it received substantial creative refinement recommendations. An alternative storyline was proposed for Pair 7 (i.e., argumentative and not argumentative) to improve understanding. For these animations, participants recommended changing the scenario significantly from a mother arguing with her child about leaving a playground to having two child characters arguing over toys instead. It was apparent that the presence of the mother figure gave the impression that the scenario was about 'obedience' rather than argumentative behavior. All refinement suggestions reported by participants in round 1 interviews are listed in Table 6.

3.1.4. First iteration results summary

Based on combined acceptability and accuracy ratings and refinement suggestions, twenty-two original items (73%) were retained and eight required refinement after the first round of interviews. Seven individual animations were retained from pool one (i.e., items 1a and b, 2a, 3b, 4a and b, and 6b) and seven required further refinement. Those requiring refinement were items 2b (not worried), 3a (sleeps poorly), 5a (disobedient at school), 5b (obedient at school), 6a (shy), 7a (argumentative), and 7b (not argumentative). From pool two participant responses, 15 individual animations were retained (i.e., items 8b, 9a and b, 10a and b, 11a and b, 12a and b, 13a and b, 14a and b, and 15a and b), and one item (8a, hyperactive) required refinement.

3.2. Second iteration results

The second round of interviews were conducted with a third and fourth pool of participants. Pool three participants ($n = 14$, 50% female, $M_{age} = 8$ years) examined five pairs of animations. Of the ten individual animations they reviewed, seven had been refined (i.e., items 2b, 3a, 5a, 5b, 6a, 7a, and 7b). Animation 8a (i.e., hyperactive) required multiple technical refinements due to the difficulty of animating some of the actions before it was acceptable and was not finalized in time for examination by pool three participants. Therefore, the refined version of this animation along with its pair (i.e., animation, 8b calm and sensible) was examined by pool four participants ($n = 7$, 42.9% female, $M_{age} = 8.00$ years, $SD = 2.20$).

3.2.1. Animation acceptability

Acceptability of the animations was again high with 100% rated ≥4.0 stars. The average satisfaction ratings out of five stars are reported in Table 7 for each animation. In participant pool three, the average star-rating across all animations was 4.5 for audio appeal, 4.6 for character appeal, 4.6 for animated action and context, and 4.8 for viewing appeal. These all increased on the ratings given in round one interviews. Combined satisfaction ratings were high across all animations with an average total satisfaction score of 4.6. The average star-rating for both items viewed by pool four participants was 4.4 for audio appeal, 5.0 for character appeal, 4.9 for animated action and context, and 5.0 for viewing appeal. Combined satisfaction ratings were high with an average total satisfaction score of 4.9. No refinements were warranted based on these satisfaction ratings.

Table 4
Iteration one: number and proportion of correct responses assessing original item accuracy.

Animated Item	Correct Understanding N (%)	Correct Identification N (%)	Correct Representation N (%)	Correct Audio N (%)	Total correct N (%)
Participant Pool 1 (N = 21)					
1a Sad – Depressed	21 (100)	20 (95.3)	21 (100)	21 (100)	83 (98.8)
1b Happy	21 (100)	20 (95.3)	21 (100)	20 (95.3)	82 (97.6)
2a Worried – Anxious	20 (95.3)	20 (95.3)	21 (100)	20 (95.3)	81 (96.4)
2b Not worried – Confident	19 (90.5)	18 (85.7)	18 (85.7)	20 (95.3)	75 (89.4)
3a Sleeps poorly	19 (90.5)	19 (90.5)	21 (100)	21 (100)	80 (95.2)
3b Sleeps well	19 (90.5)	20 (95.3)	21 (100)	21 (100)	81 (96.4)
4a Angry	20 (95.3)	19 (90.5)	21 (100)	21 (100)	81 (96.4)
4b Not angry – Impassive	17 (80.9)	21 (100)	21 (100)	21 (100)	80 (95.2)
5a Disobedient (School)	18 (85.7)	18 (85.7)	17 (80.9)	21 (100)	74 (88.1)
5b Obedient (School)	**6 (28.6)**	**10 (47.6)**	**16 (76.2)**	20 (95.3)	**52 (61.9)**
6a Shy	17 (80.9)	**16 (76.2)**	19 (90.5)	20 (95.3)	72 (85.7)
6b Not Shy – Outgoing	18 (85.7)	19 (90.5)	21 (100)	21 (100)	79 (94)
7a Argumentative	**14 (66.7)**	**14 (66.7)**	18 (85.7)	21 (100)	**67 (79.8)**
7b Not argumentative	**12 (57)**	**7 (33)**	18 (85.7)	21 (100)	**58 (69)**
Participant Pool 2 (N = 20)					
8a Hyperactive behavior	**12 (60)**	**11 (55)**	**12 (60)**	20 (100)	**55 (68.8)**
8b Calm – Sensible	17 (85)	19 (95)	19 (95)	20 (100)	75 (93.8)
9a Lonely – Alone	19 (95)	19 (95)	19 (95)	20 (100)	77 (96.3)
9b Sociable – Alone by choice	19 (95)	19 (95)	20 (100)	20 (100)	78 (97.5)
10a Bullied – Excluded	18 (90)	20 (100)	20 (100)	20 (100)	78 (97.5)
10b Not bullied – Included	20 (100)	20 (100)	20 (100)	20 (100)	80 (100)
11a Fearful – Scared	19 (95)	20 (100)	20 (100)	20 (100)	79 (98.8)
11b Not Scared – Brave	18 (90)	20 (100)	20 (100)	20 (100)	78 (97.5)
12a Disobedient (Home)	16 (80)	20 (100)	20 (100)	20 (100)	76 (95)
12b Obedient (Home)	20 (100)	20 (100)	20 (100)	20 (100)	80 (100)
13a Distracted – Inattentive	18 (90)	18 (90)	20 (100)	19 (95)	75 (93.8)
13b Focused – Pays attention	19 (95)	19 (95)	19 (95)	20 (100)	77 (96.3)
14a Physically aggressive	18 (90)	19 (95)	19 (95)	20 (100)	76 (95)
14b Kind – Peaceful	18 (90)	20 (100)	20 (100)	20 (100)	78 (97.5)
15a Physical symptoms – Feel sickly	20 (100)	20 (100)	20 (100)	20 (100)	80 (100)
15b No physical symptoms – Feel well	18 (90)	20 (100)	20 (100)	20 (100)	78 (97.5)

Note: Bold denotes items with <80% accuracy consensus targeted for refinement.

Table 5
Iteration one: items with accuracy rating variations < 80% as a function of age group level.

Item		Correct Understanding n(%)			Correct Identification n(%)			Correct Representation n(%)		
		5–6 Years	7–9 Years	10–11 Years	5–6 Years	7–9 Years	10–11 Years	5–6 Years	7–9 Years	10–11 Years
Participant Pool 1 (N = 21)		*n = 5*	*n = 8*	*n = 8*	*n = 5*	*n = 8*	*n = 8*	*n = 5*	*n = 8*	*n = 8*
2b	**Not Worried - Confident**	5(100)	**6(75)**	8(100)	5(100)	7(87.5)	**6(75)**	5(100)	**6(75)**	7(87.5)
4b	**Not Angry - Impassive**	**3(60)**	**6(75)**	8(100)	5(100)	8(100)	8(100)	5(100)	8(100)	8(100)
5a	**Disobedient (School)**	4(80)	7(87.5)	7(87.5)	4(80)	7(87.5)	7(87.5)	5(100)	8(100)	**4(50)**
5b	**Obedient (School)**	5(100)	**1(12.5)**	**0(0%)**	4(80)	**3(37.5)**	**3(37.5)**	5(100)	**5(62.5)**	**6(75)**
6a	**Shy**	5(100)	7(87.5)	**5(62.5)**	4(80)	**6(75)**	**6(75)**	5(100)	7(87.5)	7(87.5)
6b	**Not Shy Outgoing**	5(100)	7(87.5)	**6(75)**	5(100)	7(87.5)	7(87.5)	5(100)	8(100)	8(100)
7a	**Argumentative**	**3(60)**	7(87.5)	**4(50)**	4(80)	**5(62.5)**	**5(62.4)**	5(100)	**6(75)**	7(87.5)
7b	**Not Argumentative**	**2(40)**	**6(75)**	**4(50)**	**3(60)**	**1(12.5)**	**3(37.5)**	5(100)	7(87.5)	**6(75)**
Participant Pool 2 (N = 20)		*n = 5*	*n = 9*	*n = 6*	*n = 5*	*n = 9*	*n = 6*	*n = 5*	*n = 9*	*n = 6*
8a	**Hyperactive**	**2(40)**	**5(55.6)**	5(83.3)	**1(20)**	**6(66.7)**	**4(66.7)**	**1(20)**	8(88.9)	**3(50)**
8b	**Calm Sensible**	**2(40)**	9(100)	6(100)	4(80)	9(100)	6(100)	4(80)	9(100)	6(100)
10a	**Bullied Excluded**	5(100)	**7(77.8)**	6(100)	5(100)	9(100)	6(100)	5(100)	9(100)	6(100)
11b	**Not Scared Brave**	5 (100)	**7(77.8)**	6 (100)	5 (100)	9 (100)	6 (100)	5 (100)	9 (100)	6(100)
12a	**Disobedient (Home)**	4(80)	**7(77.8)**	5(83.3)	5(100)	9(100)	6(100)	5(100)	9(100)	6(100)
13a	**Distracted Inattentive**	**3(60)**	9(100)	6(100)	4(80)	8(88.9)	6(100)	5(100)	8(88.9)	6(100)

Note: Bold denotes items with <80% accuracy consensus.

3.2.2. Animation accuracy

Accuracy was again examined across the whole participant pool and within age group levels for each response category. Across the sample, all items reviewed by pool three participants reached at least 92.9% accuracy consensus ratings in each response category (i.e., 'understanding', 'identification', 'representation', and 'audio') with 60% of items achieving 100% accuracy in each category. However, there were still discrepancies within age group level one participants regarding the accuracy of items 5a and 5b (disobedient and obedient: school context), and 7a and 7b (argumentative and not argumentative).

Items 5a and 5b failed to reach 80% consensus in the 'understanding' response category for age group level one. On examination of individual responses, it was noted that this was due to one participant (25%) aged five years being unable to provide an equivalent personal example for items 5a and 5b because he did not yet attend school. Items 5a and 5b reached 100% consensus in all age group levels (including level one) in all other response categories, therefore these animations were retained. Items 7a (argumentative) and 7b (not argumentative) again failed to reach 80% consensus within the 5–6-year age group level in both 'understanding' and 'identification' categories. Half of these younger

Table 6
Iteration one: participants' suggestions for items requiring refinement.

Item		Add, remove or change animation action, sound, scenario,
2b	Not worried – Confident	make the 'ding' sound louder*; have her say "aw yeah". add a fist pump*; add a thumbs-up*; put her hands on her hips, make her wink bigger*. remove the #1 symbol from the thought bubble
3a	Sleeps poorly	add moaning*, add a big sigh*; make the yawn louder; add footstep sounds, put an angry face on the clock; put eyes on the claws, make the monsters bigger*scarier*. change the monsters to rude/mean people
5a	Disobedient (School)	make boy poke out his tongue* teacher needs to say something; say "no" out loud*; make him wear his hat on backwards and play with a ball; have the teacher yell at him; have boy say "no" and teacher say "behave"; put a toy on his desk
5b	Obedient (School)	remove the stars 'ding' sound*; make teacher clap hands to get his attention; remove the star*; nod the boys head as if saying "yes" to the teacher; have boy say "okay"; add 'ok' to bubble; remove math from the bubble. Add boy saying, "yes miss"; have teacher say "yes "when he puts his hand up.
6a	Shy	have her face away from the kids instead of having her head down girl mumbles "no thank you"; change mm mm to uh uh. remove "mm mmm" said by girl; have her picking grass; remove the book; don't cover her face as much. Make her look sadder.
7a	Argumentative	add some voices*; arguing sounds*; arguing sounds back and forth*. show more angry faces*; show them both arguing; add 'swirly' symbols above her head* and exclamation marks; start it in the sandpit*; make it two kids instead*. make it shorter; take the mum out*, have them arguing/fighting over toys*
7b	Not argumentative	add some voices*; add friendly sounds*; talking nicely sounds back and forth*. show happy faces*; start it in the sandpit*; make it two kids*. make it shorter; take the mum out*, have them arguing/fighting over toys*
8a	Hyperactive behavior	have him running* spinning around*; swing on his chair*; build a fort with books; run around his desk fast*; say "sir, sir, sir"; more fiddling around*; throw paper; run around in a circle and act sillier*; add silly head movement; more active*, more silly type stuff with a cheeky face; add pencil sounds, book thuds.

Note. * = suggestion made multiple times.

Table 7
Iteration two: average satisfaction ratings (out of 5) per item.

Animated Item	Audio Appeal	Character appeal	Animated Action Context	Viewing Appeal	Total Average Satisfaction
Pool 3 (N = 14)					
2a	4.1	4.5	4.6	4.9	4.5
2b*	4.0	4.2	4.1	4.6	4.3
3a*	4.6	5.0	4.8	4.9	4.8
3b	4.6	4.9	4.8	4.7	4.8
5a*	4.6	4.7	4.9	5.0	4.8
5b*	5.0	4.5	4.6	4.9	4.8
6a*	4.8	4.9	4.9	5.0	4.9
6b	4.5	4.6	4.7	4.8	4.6
7a*	4.4	4.7	4.6	4.9	4.6
7b*	4.4	4.6	4.7	4.9	4.6
Pool 4 (N = 7)					
8a*	4.4	5.0	5.0	5.0	4.9
8b	4.3	5.0	4.7	5.0	4.8

* Refined animation.

participants could not provide a personal example equivalent to the target argumentative behavior in animation 7a; 25% could not for the contrasting animation (7b); and 25% could not identify (i.e., label) either animation. For this reason, items 7a and 7b were targeted for refinement to improve their accuracy for younger ages.

The second iteration of animation 8a (hyperactive) reviewed by pool four participants attained an accuracy rating of 100% (for all response categories) both across the whole sample and within age group levels. Table 8 identifies accuracy ratings across the whole sample (5–11-years) for each animation accuracy response category for all items examined during Iteration Two.

3.2.3. Refinement feedback for identified animations

Refinement suggestions offered by participants to improve the understanding of animations 7a and 7b predominantly related to adding dialogue between characters, increasing the length of the animation, and increasing the intensity of the demonstrated 'argument'. No participants proposed an alternative animation story or scenario that they thought would improve either animation. Specific suggestions for animation 7a (argumentative) were to "make the whole thing longer", "make the girls car pink and boys car blue", "make the boy talk", "remove high pitch talking and have real words", "add a bit more arguing back to each other", and "make them angrier". "Use real words" and "make them more friendly" were the only suggestions given for animation 7b (not argumentative) in this round.

3.2.4. Second iteration results summary

After the second round of interviews were completed and combined acceptability and accuracy rating and refinement suggestions analyzed, six items out of eight were retained (i.e., items 2b, 3a, 5a, 5b, and 6a). Though animations 7a (argumentative) and 7b (not argumentative) had high accuracy ratings and high acceptability overall, these two animations were identified for further refinement and examination in a third interview round. Firstly, because they had been substantially refined after round one and thus the researchers deemed it would be prudent to trial these items again with another pool of participants. Secondly, we wanted to incorporate as many improvement suggestions required from round two feedback to increase the understanding and identification accuracy ratings of younger participants. Therefore, excluding the addition of "real word" dialogue, all participant suggestions were incorporated into third versions of these animations.

3.3. Third and final iteration results

Participant pool four ($n = 7$, 42.9% female, $M_{age} = 8.00$ years, $SD = 2.20$) examined refined items 7a (argumentative) and 7b (not argumentative) in a third and final round of interviews.

3.3.1. Combined animation acceptability, animation accuracy and refinement suggestions

The acceptability of both animations was high. Item 7a and 7b rated 4.9 and 4.8 out of 5 respectively for character appeal. Both rated 5 out of 5 in all other categories: audio and viewing appeal, animated action context and total average satisfaction ratings in this final version. Both items reached accuracy ratings of 100% in 'understanding', 'identification', 'representative', and 'audio' accuracy categories, across the whole sample and within age group levels and no refinement suggestions were provided.

3.3.2. Final iteration results summary

After the third round of interviews were completed and responses analyzed, items 7a and 7b were retained. Thus, by the third and final iteration, all 15 item pairs (30 individual animations) were retained and deemed to be accurate and acceptable to participants.

3.4. Overall results

In total 25 individual animations (83%) achieved minimum accuracy requirements (at least 80% correct) in their original prototype format in every category (15 of which achieved >95% accuracy). Five failed to reached minimum accuracy requirements with ratings ranging between

Table 8
Iteration two: number and proportion of correct responses assessing accuracy of refined items.

Animated Item		Correct Understanding N (%)	Correct Identification N (%)	Correct Representation N (%)	Correct Audio N (%)	Total correct N (%)
Participant Pool 3 (N = 14)						
2b	Not worried – Confident	14 (100)	14 (100)	14 (100)	14 (100)	56 (100)
3a	Sleeps poorly	14 (100)	14 (100)	14 (100)	14 (100)	56 (100)
5a	Disobedient (School)	13 (92.9)	14 (100)	14 (100)	14 (100)	55 (98.2)
5b	Obedient (School)	13 (92.9)	14 (100)	14 (100)	14 (100)	55 (98.2)
6a	Shy	14 (100)	14 (100)	14 (100)	14 (100)	56 (100)
7a	Argumentative	12 (92.9)	13 (92.9)	14 (100)	14 (100)	53 (94.6)
7b	Not argumentative	13 (92.9)	13 (92.9)	14 (100)	14 (100)	54 (96.4)
Participant Pool 4 (N = 7)						
8a	Hyperactive behavior	7 (100)	7 (100)	7 (100)	7 (100)	28 (100)

28.6 and 76.2%. These five items were targeted for refinement along with a further three animations that received multiple refinement suggestions (despite high acceptability and accuracy ratings. Of these eight items, six (20%) required at least one round of refinements and two (7%) required two rounds of refinements before acceptability and accuracy ratings were greater than 80%. Still images from each of the final 30 animated items are shown in Fig. 2. A flowchart outlining the complete iterative item development and refinement process is presented in Fig. 3.

4. Discussion

In the current study we sought to qualitatively confirm digitally animated assessment item content for inclusion in the ICDS with the target audience of the measure: that is children aged between five and 11 years. A major focus of this research was the utilization of participatory methodologies that included children as co-designers of the animated item content. First to determine how acceptable the digital animations were to them and second to assess the accuracy of the content of the animations and increase accuracy where required via iterative refinement processes to promote optimal recognition and understanding.

4.1. Principal findings

Our study revealed two key findings. First, that digitally animated items were acceptable to all participants in this study. Satisfaction

Fig. 2. Still images from each of the target and contrasting animated items (N = 30).

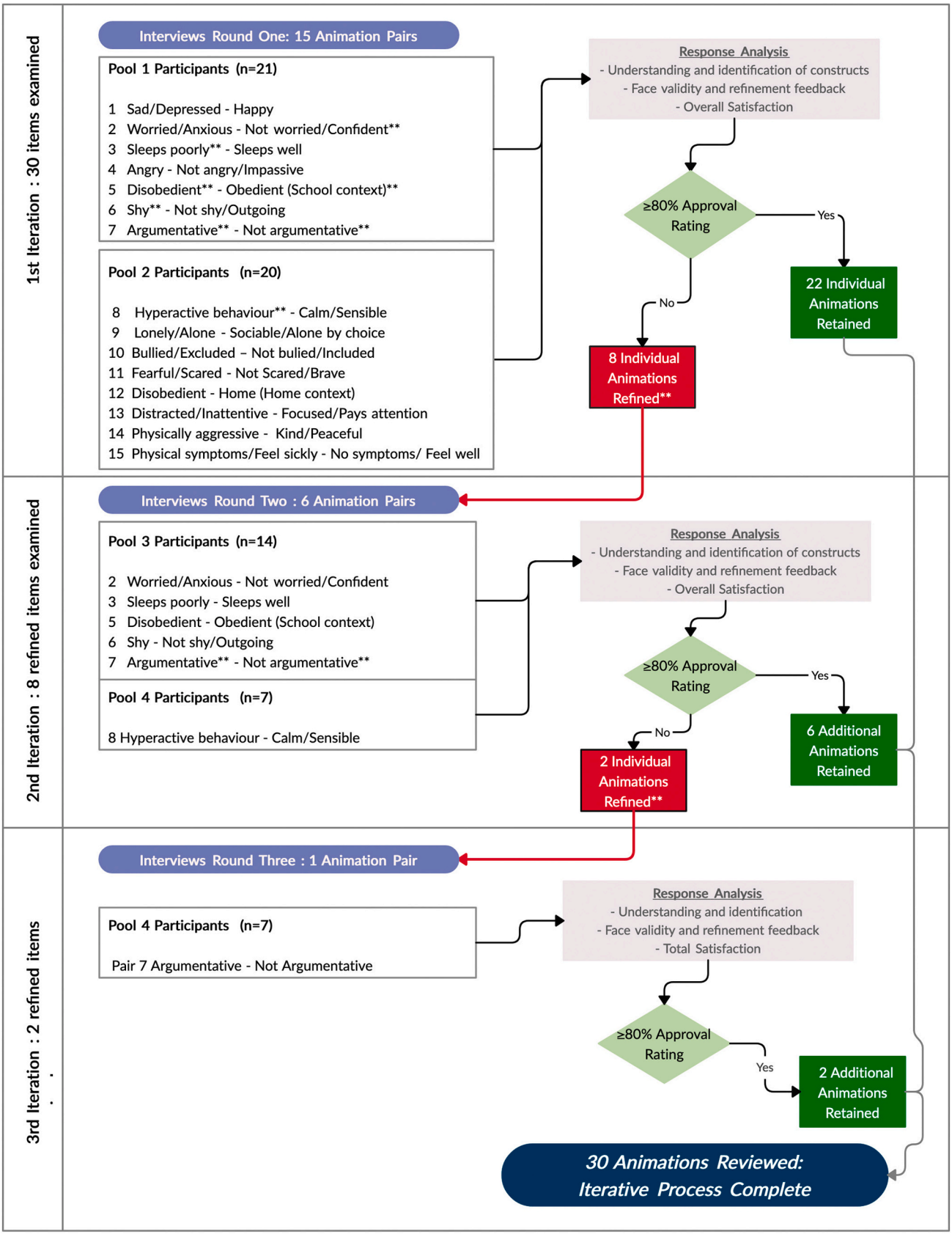

Fig. 3. Iterative item development and refinement process and results
Note. ** indicates the specific animations that required refinement and were moved to a following interview round for evaluation and potential refinement.

ratings were high, even for those items that were later identified by children as requiring refinement. This was evident across audio, character, and scenario appeal and the overall viewing appeal of each animation. Satisfaction ratings were further improved with the second and third iterations and averaged 4.8 out of 5. Thus, we conclude that animated item content depicting emotional and behavioural issues is an acceptable format to children aged 5 through 11 years.

These results echo previous efforts at digitizing measures for children and including pictorial representations such as the Dominic-R (Valla et al., 2000) and MAAC (Manassis et al., 2013) which demonstrated that children showed favorable opinions of digital assessments that incorporated visual images. However, these instruments tended to utilize only static images and characters to deliver assessment items and still relied on written words or a professional to read assessment questions. Importantly, no other instrument has assessed full animations as a means of demonstrating emotional or behavioural constructs. The results of this study, along with the high acceptability demonstrated in our original feasibility study (March et al., 2018) show that this approach is highly acceptable to children. It is worth noting that the ICDS instrument has been developed following a co-design and generative participatory design process throughout all stages (March et al., 2018; Zieschank et al., 2020), and it seems likely that this has contributed to the overall high appeal of the animations.

Second, through a generative participatory design and iterative refinement process, this study was able to produce 15 pairs of contrasting animations depicting common emotional and behavioural constructs, for which children aged 5–11 years showed excellent understanding and identification. There were some challenges in understanding and identification accuracy in early animation iterations, typically for more complex constructs. We were able to use iterative co-design processes to effectively advance items that were originally difficult for young children to understand, to a point where accuracy was equivalent across age groups. This highlights the utility and importance of involving children of all ages and persisting with iterative development until accuracy and understanding is achieved.

Importantly, the findings from this study revealed that a lack of verbal labelling ability (e.g., being able to accurately label 'worry' or 'argumentative') did not necessarily indicate misunderstanding by participants. Some children were able to demonstrate accurate internal representations of the animated emotional and behavioural constructs more easily than they were able to produce accurate lexical labels for them. That is, children could describe a similar scenario in which they or others had felt or behaved like the depicted construct even when they could not name it. This provides further support for the notion that focusing on written or verbal labels of emotions and behaviours may be difficult for children with less advanced vocabularies, but that through visual stimuli, children as young as five appear able to understand the same constructs. For example, in written question items such as "I often feel hyperactive," being unable to understand the target word "hyperactive" would lead to an inability to answer the correction appropriately. However, in the case of the ICDS, the animated item would provide enough context for the child to recognize the intended construct (even if they couldn't label it), and therefore elicit an appropriate response.

Thus, the results of this study show the potential of animated scenarios demonstrating dynamic facial expressions and behaviours to overcome barriers typical to question and answer instruments that require the child to read, understand and respond to sophisticated statements about their wellbeing (Widen and Russell, 2010, 2015). Such findings provide further evidence against the notion that younger children are unable to reliably self-report on their own mental health and support a growing body of evidence which shows children's capacity to provide accurate subjective reflections on their emotions and behaviours (Cree et al., 2002; Hudziak et al., 2007; Kirk, 2007; Riley, 2004).

4.2. Strengths and limitations

The primary strength of this study is its use of generative participatory design methodology with children (< 11 years) to improve the accuracy and acceptability of each animation. Specifically, a co-design approach was utilized across all stages of item and animation development and refinement in this study, using a sample of children who were at the target age for the ICDS instrument. To accommodate for the methodological challenges of conducting participatory design research with young children, the study implemented measures and rating systems in ways that were familiar to children and easy to understand. For example, we utilized star ratings and image-based questions to obtain data on acceptability and content validity. Importantly, the study contextualized the research aims to the needs and ability of the target group and utilized multiple measures of accuracy and outcome (e.g., internal representation, lexical labelling), to tap into children's ability to understand the intended constructs.

Notwithstanding these strengths, there were also some limitations. Given the aim of this study was to examine acceptability of the item format and content validity of the animations prior to selection for the ICDS instrument, we prioritized data collection via in-depth interviews, which allowed child participants to engage with the animations comprehensively and provide lengthy answers to questions. A consequence of this was that the sample size ($N = 62$ children) precluded in-depth comparisons of responses between ages and genders. For this study, which focused on early item development and establishing item accuracy and acceptability, this sample size was sufficient. However, full psychometric validation of the prospective measure is still required with much larger samples. An additional limitation is that our sample was relatively homogenous in that participants were predominantly white Australians of middle to high socio-economic status. It would be beneficial to recruit a more diverse sample in future research to identify whether certain cultural or sociodemographic factors influence item understanding and acceptability. Further, although neurodiversity and clinical status was not formally assessed, it is likely that this sample was not comprised of children with clinical level difficulties or developmental delays given the community sample recruitment strategy. These groups will require targeted recruitment in future research to determine applicability of the animated item content.

4.3. Implications

The findings of this research highlight that children as young as five have the ability to understand animated depictions of emotions and behaviours that serve as exemplars of emotional and behavioural distress and can apply them to themselves (internal representation). This has potential implications for the way such constructs are assessed (e.g., screening instruments) and clinical practice. As demonstrated here, animations potentially provide a novel and useful mechanism for obtaining accurate self-report data from young children. If the ICDS can be validated psychometrically, it will provide new opportunities for self-report assessment of children under 11 years of age that can inform our understanding of distress and wellbeing, from the perspective of the child as an adjunct to proxy report. Given the well-documented discrepancies between parent-child agreement on paper and pencil type measures, the findings of this study support a promising new approach to obtaining multi-informant, multi-method data on children's wellbeing.

4.4. Conclusion

The perspectives of children under the age of 11 have typically been neglected in assessments designed to provide self-report of child emotions, behaviours, or general wellbeing. The results of this study provide support for the notion that new digital technologies such as dynamic animations may be able to overcome some of the potential barriers to

conducting self-rated assessment with younger children. In this study, children aged 5 to 11 years were able to accurately identify and understand complex emotions and behaviours via engaging digital animated items. Overall, this study highlights the general willingness of children to engage with (and the appeal of) digital animations designed to assess distress or mental health. The study also shows the potential of animated items to accurately convey depictions of emotional and behavioural constructs key to childhood disorders, especially when co-designed and refined through an iterative process.

Funding

This research was supported by an Australian Government Research Training Scheme scholarship.

Declaration of competing interest

The authors have no conflicts of interest to declare.

References

Achenbach, T., & Rescorla, L. (2003). Achenbach Assessment - School age. In *Manual for the ASEBA School-Age Forms & Profiles* (pp. 99–107). Research Center for Children, Youth, and Families.

Achenbach, T.M., McConaughy, S.H., Ivanova, M.Y., Rescorla, L.A., 2011. Manual for the ASEBA Brief Problem Monitor (BPM). University of Vermont, Research Center for Children Youth and Families. https://aseba.org/school-age-bpm/.

Bergeron, L., Berthiaume, C., St-Georges, M., Piché, G., Smolla, N., 2013. Reliability, validity, and clinical use of the Dominic Interactive: A DSM-based, self-report screen for school-aged children. Can. J. Psychiatr. 58 (8), 466–475. https://doi.org/10.1177/070674371305800805.

Caspi, A., Henry, B., McGee, R.O., Moffitt, T.E., Silva, P.A., 1995. Temperamental origins of child and adolescent behavior problems: from age three to age fifteen. Child Dev. 66 (1), 55–68. https://doi.org/10.1111/j.1467-8624.1995.tb00855.x.

Children''s Health Queensland. (2018). *Children's Health and Wellbeing Services Plan 2018–2028: Children's Health Queensland Hospital and Health Service*.

Cockburn, A., 2008. Using both incremental and iterative development. CrossTalk 21 (5), 27–30.

Cree, V.E., Kay, H., Tisdall, K., 2002. Research with children: sharing the dilemmas. Child and Family Social Work 7 (1), 47–56. https://doi.org/10.1046/j.1365-2206.2002.00223.x.

Cugelman, B., 2013. Gamification: what it is and why it matters to digital health behavior change developers. J. Med. Internet Res. 15 (12) https://doi.org/10.2196/games.3139.

Darbyshire, P., MacDougall, C., Schiller, W., 2005. Multiple methods in qualitative research with children: more insight or just more? Qual. Res. 5 (4), 417–436. https://doi.org/10.1177/1468794105056921.

Edwards, J., Parson, J., O'Brien, W., O'Brien, W., O'Brien, W., 2016. Child play therapists' understanding and application of the United Nations convention on the rights of the child: A narrative analysis. International Journal of Play Therapy 25 (3), 133–145. https://doi.org/10.1037/pla0000029.

Eiser, C., Morse, R., 2001. Can parents rate their child's health related quality of life? Results of a systematic review. Qual. Life Res. 10 (4), 347–357.

Farcic, V., 2014. Software development models: iterative and incremental development | Technology conversations. Technology Conversations 1.

Goodman, R., 1997. The strengths and difficulties questionnaire: a research note. Journal of Child Psychology and Psychiatry and Allied Disciplines 38 (5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x.

Greene, S., Hogan, D. (Eds.), 2005. Researching Children's Experience. SAGE Publications Ltd. https://doi.org/10.4135/9781849209823

Hudziak, J., Achenbach, T., Althoff, R., Pine, D., 2007. A dimensional approach to developmental psychopathology. Int. J. Methods Psychiatr. Res. 16 (S1), S16–S23. https://doi.org/10.1002/mpr.

Jacka, F.N., Reavley, N.J., 2014. Prevention of mental disorders: evidence, challenges and opportunities. BMC Med. 12 (1), 1–3. https://doi.org/10.1186/1741-7015-12-75.

Jardine, J., Glinianaia, S.V., McConachie, H., Embleton, N.D., Rankin, J., 2014. Self-reported quality of life of young children with conditions from early infancy: a systematic review. Pediatrics 134 (4), e1129–e1148. https://doi.org/10.1542/peds.2014-0352.

Johnson, S.E., Lawrence, D., Hafekost, J., Saw, S., Buckingham, W.J., Sawyer, M., Ainley, J., Zubrick, S.R., 2016. Service use by Australian children for emotional and behavioural problems: findings from the second Australian Child and Adolescent Survey of Mental Health and Wellbeing. Aust. N. Z. J. Psychiatry 50 (9), 887–898. https://doi.org/10.1177/0004867415622562.

Kirk, S., 2007. Methodological and ethical issues in conducting qualitative research with children and young people: A literature review. Int. J. Nurs. Stud. 44 (7), 1250–1260. https://doi.org/10.1016/j.ijnurstu.2006.08.015.

Lawrence, D., Hafekost, J., Johnson, S.E., Saw, S., Buckingham, W.J., Sawyer, M.G., Ainley, J., Zubrick, S.R., 2016. Key findings from the second Australian child and adolescent survey of mental health and wellbeing. Aust. N. Z. J. Psychiatry 50 (9), 876–886. https://doi.org/10.1177/0004867415617836.

Linares Scott, T.J., Short, E.J., Singer, L.T., Russ, S.W., Minnes, S., 2006. Psychometric properties of the Dominic interactive assessment. Assessment 13 (1), 16–26. https://doi.org/10.1177/1073191105284843.

Manassis, K., Mendlowitz, S., Kreindler, D., Lumsden, C., Sharpe, J., Simon, M.D., Woolridge, N., Monga, S., Adler-Nevo, G., 2009. Mood assessment via animated characters: a novel instrument to evaluate feelings in young children with anxiety disorders. Journal of Clinical Child and Adolescent Psychology 38 (3), 380–389. https://doi.org/10.1080/15374410902851655.

Manassis, K., Mendlowitz, S., Dupuis, A., Kreindler, D., Lumsden, C., Monga, S., Guberman, C., 2013. Mood assessment via animated characters: an instrument to access and evaluate emotions in young children. Open Journal of Psychiatry 03 (01), 149–157. https://doi.org/10.4236/ojpsych.2013.31a010.

March, S., Day, J., Zieschank, K., Ireland, M., 2018. The interactive child distress screener: development and preliminary feasibility testing. JMIR MHealth and UHealth 6 (4), e90. https://doi.org/10.2196/mhealth.9456.

Marsac, M.L., Winston, F.K., Hildenbrand, A.K., Kohser, K.L., March, S., Kenardy, J., Kassam-Adams, N., 2015. Systematic, theoretically grounded development and feasibility testing of an innovative, preventive web-based game for children exposed to acute trauma. Clinical Practice in Pediatric Psychology 3 (1), 12–24. https://doi.org/10.1037/cpp0000080.

Mellor, D., Moore, K.A., 2014. The use of Likert scales with children. J. Pediatr. Psychol. 39 (3), 369–379. https://doi.org/10.1093/jpepsy/jst079.

Moreira, H., Carona, C., Silva, N., Frontini, R., Bullinger, M., Canavarro, M.C., 2013. Psychological and quality of life outcomes in pediatric populations: a parent-child perspective. J. Pediatr. 163 (5), 1471–1478. https://doi.org/10.1016/j.jpeds.2013.06.028.

Mummah, S.A., Robinson, T.N., King, A.C., Gardner, C.D., Sutton, S., 2016. IDEAS (Integrate, Design, Assess, and Share): A framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J. Med. Internet Res. 18 (12), e317 https://doi.org/10.2196/jmir.5927.

Newton, A.S., Wozney, L., Bagnell, A., Fitzpatrick, E., Curtis, S., Jabbour, M., Johnson, D., Rosychuk, R.J., Young, M., Ohinmaa, A., Joyce, A., McGrath, P., 2016. Increasing access to mental health care with breathe, an internet-based program for anxious adolescents: study protocol for a pilot randomized controlled trial. JMIR Research Protocols 5 (1), e18. https://doi.org/10.2196/resprot.4428.

Patrick, K., Hekler, E.B., Estrin, D., Mohr, D.C., Riper, H., Crane, D., Godino, J., Riley, W. T., 2016. The pace of technologic change: implications for digital health behavior intervention research. Am. J. Prev. Med. 51 (5), 816–824. https://doi.org/10.1016/j.amepre.2016.05.001.

Riley, A.W., 2004. Evidence that school-age children can self-report on their health. Ambul. Pediatr. 4 (4), 371–376. https://doi.org/10.1367/A03-178R.1.

Royal Australian and New Zealand College of Psychiatrists (RANZCP). (2017). *Advocating for mental health resources commensurate with the burden of disease: 2017–18 Pre-Budget submission report* (Issue January).

Sawyer, M.G., Arney, F.M., Baghurst, P.A., Clark, J.J., Graetz, B.W., Kosky, R.J., Nurcombe, B., Patton, G.C., Prior, M.R., Raphael, B., Rey, J.M., Whaites, L.C., Zubrick, S.R., 2001. The mental health of young people in Australia: key findings from the child and adolescent component of the national survey of mental health and well-being. Aust. N. Z. J. Psychiatry 35 (6), 806–814. https://doi.org/10.1046/j.1440-1614.2001.00964.x.

Soffer, M., Ben-Arieh, A., 2014. School-aged children as sources of information about their lives. In: Melton, G., Ben-Arieh, A., Cashmore, J., Goodman, G., Worley, N. (Eds.), The SAGE Handbook of Child Research. SAGE Publications Ltd. https://doi.org/10.4135/9781446294758

Stålberg, A., Sandberg, A., Söderbäck, M., Larsson, T., 2016. The child's perspective as a guiding principle: Young children as co-designers in the design of an interactive application meant to facilitate participation in healthcare situations. J. Biomed. Inform. 61, 149–158. https://doi.org/10.1016/j.jbi.2016.03.024.

Stoyanov, S.R., Hides, L., Kavanagh, D.J., Sanders, D., Cockshaw, W., Mani, M., 2016. A recommended process for development and evaluation of eTools for mental health and wellbeing. In: Menzies, R.G., Kyrios, M., Kazantzis, N. (Eds.), Innovations and Future Directions in the Behavioural and Cognitive Therapies. Australian Academic Press, Issue June.

Ten Brummelaar, M.D.C., Kalverboer, M.E., Harder, A.T., Post, W.J., Zijlstra, A.E., Knorth, E.J., 2014. The Best Interest of the Child Self-Report questionnaire (BIC-S): results of a participatory development process. Child Indic. Res. 7 (3), 569–588. https://doi.org/10.1007/s12187-013-9225-3.

The Royal Australian and New Zealand College of Psychiatrists. (2010). *The Prevention and Early Intervention of Mental Illness in Infants, Children and Adolescents*. *October*, 63–65.

Truman, J., Robinson, K., Evans, A.L., Smith, D., Cunningham, L., Millward, R., Minnis, H., 2003. The strengths and difficulties questionnaire: A pilot study of a new computer version of the self-report scale. Eur. Child Adolesc. Psychiatry 12 (1), 9–14. https://doi.org/10.1007/s00787-003-0303-9.

Valla, J.-P., Bergeron, L., Smolla, N., 2000. The Dominic-R: A pictorial interview for 6- to 11-year-old children. J. Am. Acad. Child Adolesc. Psychiatry 39 (1), 85–93. https://doi.org/10.1097/00004583-200001000-00020.

Vandekerckhove, P., De Mul, Ma., Bramer, W. M., & De Bont, A. A. (2020). Generative participatory design methodology to develop electronic health interventions: systematic literature review. J. Med. Internet Res., 22(4), 1–18. doi:https://doi.org/10.2196/13780.

Widen, S.C., Russell, J.A., 2010. Children's scripts for social emotions: causes and consequences are more central than are facial expressions. Br. J. Dev. Psychol. 28 (3), 565–581. https://doi.org/10.1348/026151009X457550d.

Widen, S.C., Russell, J.A., 2015. Do dynamic facial expressions convey emotions to children better than do static ones? J. Cogn. Dev. 16 (5), 802–811. https://doi.org/10.1080/15248372.2014.916295.

Wood, B.J., McDaniel, T., 2020. A preliminary investigation of universal mental health screening practices in schools. Child Youth Serv. Rev. 112 (March), 104943. https://doi.org/10.1016/j.childyouth.2020.104943.

Wrobel, A. (2019). Young children's use of digital technologies : Risks and opportunities for early childhood development [CoLab Evidence Report]. Retrieved from *Https://Colab.Telethonkids.Org.Au/Resources/*.https://colab.telethonkids.org.au/resources/.

Yarosh, S., Schueller, S.M., 2017. "Happiness inventors": informing positive computing technologies through participatory design with children. J. Med. Internet Res. 19 (1) https://doi.org/10.2196/jmir.6822.

Zieschank, K., Machin, T., Day, J., Ireland, M., & March, S. (2020). [Manuscript submitted for publication] Understanding Children's Perspectives on Emotions and Behaviours to Develop a Child-Reported Screening Instrument. School of Psychology and Counselling and Centre for Health Research, University of Southern Queensland.