International Journal of Nursing Studies Advances 3 (2021) 100038

Contents lists available at ScienceDirect

International Journal of Nursing Studies Advances

journal homepage: www.sciencedirect.com/journal/international-journal-of-nursing-studies-advances

How to screen for social withdrawal in primary care: An evaluation of the alarm distress baby scale using item response theory

Ida Egmose [a,*], Johanne Smith-Nielsen [a], Theis Lange [b], Maria Stougaard [a], Anne C. Stuart [a], Antoine Guedeney [c], Mette Skovgaard Væver [a]

[a] Center for Early Intervention and Family Studies, Department of Psychology, University of Copenhagen, Copenhagen, Denmark
[b] Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[c] Université de Paris, et Hôpital Bichat Claude Bernard APHP, Paris, France

ARTICLE INFO

Keywords:
Alarm distress baby scale
ADBB
Social withdrawal
Early detection
Construct validity
Item response theory

ABSTRACT

Background: Early identification of infants at-risk is imperative for proper referral to intervention programs. The Alarm Distress Baby Scale (ADBB) is an eight-item observer-rated screening tool detecting social withdrawal in infants. Previously, a shortened five-item version of the scale (m-ADBB) has been proposed. To date, few studies have examined the validity of the two scales, and no studies have examined the validity of the ADBB after implementation as a universal screening tool in primary care.

Objective: The aim of this study is to use Item Response Theory (IRT) to examine the construct validity of the ADBB when used by public health visitors in primary care.

Methods: Participants were 24,752 infants (aged: 2-12.9 months) screened by public health visitors using the ADBB. Screenings were categorized into three waves according to the infant's age at the screening time (2-3.9 months, 4-7.9 months, and 8-12.9 months). Analyses were conducted separately on each wave. We checked IRT assumptions: (a) Unidimensionality, (b) Monotonicity, (c) Local independence, and (d) No DIF in relation to infant sex and gestational age. The 2PLM was used to assess model fit and estimate model parameters.

Results: Items fulfilled assumptions regarding unidimensionality, monotonicity, and no clinical and significant DIF. Local independence was not present for all items (i.e. 2, 7, and 8). The items showed moderate to good discriminatory abilities (alpha values $\geq$ 1.11) and discriminated best above average levels of social withdrawal (theta values $\geq$ 1.33). Items 7 and 8 showed nearly identical ICC suggesting that the two items discriminate equally well at the same level of social withdrawal. In addition, items 4 and 6 discriminated best at very high levels of social withdrawal, which might be of limited interest for screening purposes. Finally, the items showed similar patterns in terms of discrimination and location parameters across the three waves.

Conclusions: The ADBB shows several psychometric strengths when used by public health visitors in primary care, and the items show good discriminatory abilities at the levels of social withdrawal of interest for screening purposes. Yet, the results also suggest that for first-line screening, the validity of the scale might be improved with the removal of items 4, 6, and 8 as suggested in

* Corresponding author at: Department of Psychology, University of Copenhagen, Copenhagen, Denmark.
E-mail address: i@egmose.dk (I. Egmose).

https://doi.org/10.1016/j.ijnsa.2021.100038
Received 29 December 2020; Received in revised form 8 July 2021; Accepted 12 July 2021
Available online 15 July 2021

the m-ADBB. However, before recommending implementation of the m-ADBB, studies comparing the criterion-related validity of the two scales are needed.

What is already known about this topic?

- To detect infants at-risk, formal screening programs using standardized tools are more effective than informal developmental surveillance.
- The Alarm Distress Baby Scale (ADBB) is a screening tool for detecting social withdrawal in infants aged 2-24 months, and the m-ADBB is a shortened version of the scale.
- No previous studies have examined the validity of the ADBB when used as a universal screening tool in primary care.

What this paper adds

- The ADBB shows several psychometric strengths when used as a universal screening tool by health visitors in primary care.
- The results suggest that the m-ADBB might be more appropriate for first-line screening.

1. Background

Socioemotional and behavioral problems in infancy often persist into preschool and school age (Briggs-Gowan et al., 2006), and intervention becomes more difficult as problems in infancy tend to become more complex and severe with development (Shonkoff and Phillips, 2000). To ensure referral to early intervention, studies suggest that formal screening programs, which are based on the use of standardized tools, are more effective than more informal developmental surveillance, which is based on the clinician's impression of the infant, parental concern, and input from other professionals (Briggs-Gowan et al., 2006; Guevara et al., 2013). Thus, we need well-validated screening tools, which are feasible in primary care. In this study, we examine the psychometric properties of the Alarm Distress Baby scale (ADBB; Guedeney and Fermanian, 2001) when used in a real-life primary care setting as a universal screening tool for detecting early social withdrawal in infancy. The psychometric properties are examined using Item Response Theory (IRT) on a dataset including screening results from between 11,500 to 14,500 infants aged 2-12 months screened during routine home-visits by public health visitors.

Despite knowledge about the effectiveness of universal screening programs, it is often difficult to implement universal screening tools in primary care, and several studies report that professionals often do not adhere to screening guidelines resulting in very low screening prevalence rates (e.g. Sand, 2005). In a previous study examining the implementation of the ADBB as a universal screening tool in a subsample of the health visitors from the current study, results showed that acceptable screening prevalence rates (79%) were obtained within 12 months after training (Smith-Nielsen et al., 2018). Thus, the ADBB seems to be feasible as a universal screening tool in Danish primary care. When screening tools are used in a busy real-world setting their validity might be compromised. Therefore, it is important to follow-up on the implementation of the ADBB with an examination of the validity of the instrument after its implementation in primary care.

The ADBB detects social withdrawal in infants aged 2-24 months. It is an observer-rated screening tool assessing social behavior across eight items: (1) Facial expressions, (2) Eye contact, (3) General level of activity, (4) Self-stimulating gestures, (5) Vocalizations, (6) Briskness of response to stimulation, (7) Relationship, and (8) Attraction (Guedeney and Fermanian, 2001). Infant social withdrawal is characterized by less frequent eye contact, fewer emotional displays, and less vocalizations, a decreased level of activity, and possibly by increased self-stimulation and delayed reaction time (Guedeney et al., 2013). Social withdrawal within a certain range is a normal part of caregiver-infant interactions allowing the infant to self-regulate (Brazelton et al., 1974; Field, 1981). In contrast, sustained social withdrawal may have adverse effects on child development, since this behavior limits the infant's access to the social learning environment (Guedeney et al., 2013; Smith-Nielsen et al., 2019). Indeed, studies have shown that social withdrawal in infancy is associated with less optimal outcomes within several developmental domains, such as emotional and behavioral problems (Guedeney et al., 2014; Zhou et al., 2020), poorer cognitive (Guedeney et al., 2017; Milne et al., 2009; Smith-Nielsen et al., 2019), and language development (Guedeney et al., 2016; Milne et al., 2009).

The ADBB is, to the best of our knowledge, the only observer-rated tool for time-efficient screening of early psychosocial difficulties. Typically, such difficulties are assessed using parent-reported questionnaires or comprehensive and time-intensive observational coding procedures (Bagner et al., 2012). While parent-report constitutes a time-efficient way of obtaining knowledge about the child in a number of situations, which would be very time-consuming for the professional to observe, there are also several disadvantages related to the use of parent-report. First, parents may not have the knowledge and expertise required to detect all types of infant psychosocial difficulties, especially in the case of internalizing symptoms that may be easily overlooked (Groh et al., 2012). Second, parents and observers do not necessarily agree on ratings of the infant's behavior. Studies show that parents often do not agree or show minimal agreement with observers when rating their own infant's temperament, even when parents and observers rate the same situations (Seifer et al., 2004; Stifter et al., 2008). Given these issues, it is important to supplement parent-report with observer-rated screening tools.

Previous studies have assessed the psychometric properties of the ADBB using Classical Test Theory (CTT). The studies show that

the ADBB demonstrates good inter-rater reliability, internal consistency and test-retest reliability (for a review, see Guedeney et al., 2013). Construct validity of the ADBB has been tested using factor analysis (Assumpção Jr. et al., 2002; Facuri-Lopes et al., 2008; Guedeney and Fermanian, 2001; Guedeney et al., 2013; Matthey et al., 2005). Apart from one study, which found four factors (Assumpção Jr. et al., 2002), the studies find that the ADBB consists of an interpersonal and a temperamental/non-interpersonal factor (Facuri-Lopes et al., 2008; Guedeney and Fermanian, 2001; Guedeney et al., 2013; Matthey et al., 2005). In addition, the majority of the studies find that item 4 (Self-stimulating gestures) does not load on any of the two factors, and in two of the studies, item 4 is considered a third factor (Facuri-Lopes et al., 2008; Guedeney et al., 2013). The studies differ in relation to the specific items included in the two main factors. In this respect, it should be noted that most of the studies have relatively small sample sizes ($N \leq 122$), which, together with the few measured variables in each factor (≤ 4), might have affected the stability of the factors (Kyriazos, 2018). The most recent factor analysis was conducted on a large sample of 640 infants, and results from this study confirmed the three-factor solution consisting of (a) a major interpersonal factor (explaining 37.3% of the variance) comprised of item 2 (Eye contact), item 7 (Relationship), and item 8 (Attraction), (b) a minor temperamental/non-interpersonal factor (explaining 16.3% of the variance) comprised of item 1 (Facial expressions), item 3 (General level of activity), item 5 (Vocalizations), and item 6 (Briskness of response to stimulation), and (c) an additional minor factor (explaining 13.3% of the variance) comprised of item 4 (Self-stimulating gestures) (Guedeney et al., 2013).

In 2013, a shortened and modified version of the ADBB (m-ADBB) was developed by Matthey and colleagues based on their work with the ADBB in Australia (Matthey et al., 2013, 2005). The m-ADBB is scored on a 3-point scale (rated as *Satisfactory, Possible Problem*, and *Definite Problem*) and is comprised of the following five items from the original ADBB: Item 1 (Facial expressions), item 2 (Eye contact), item 3 (General level of activity), item 5 (Vocalizations), and item 7 (Relationship). In the m-ADBB, item 4 (Self--stimulating gestures) was removed because the item showed poor interrater reliability; item 6 (Briskness of response to stimulation) was removed because the item showed high correlations with item 3 (General level of activity); and item 8 (Attraction) was removed due to high correlations with six of the other items. Apart from the original validation study (Matthey et al., 2013), two recent studies have compared the psychometric properties of the m-ADBB and the full ADBB (Pérez-Martínez et al., 2020; Ulak et al., 2020). Ulak and colleagues (2020) demonstrated that the m-ADBB showed better interrater reliability than the full ADBB, whereas the full ADBB showed better internal consistency than the m-ADBB. In contrast, Pérez-Martínez and colleagues (2020) found that both the full and the m-ADBB showed acceptable internal consistency when used to assess social withdrawal in infants with Cleft Lip and Palate malformation (CL/P) at 4 and 12 months. Overall, research comparing the ADBB and the m-ADBB has mainly focused on reliability (interrater reliability and internal consistency). While reliability is necessary to establish validity, it is not sufficient (Tavakol and Dennick, 2011). Therefore, studies focusing on how removal of items 4, 6, and 8 affects the validity of the scale are needed.

To extend the existing research on the construct validity of the ADBB, in this study, we draw on methods from Item Response Theory (IRT). IRT methods complement CTT methods by providing more in-depth assessment of the performance and validity of each item on the scale. Thus, an IRT analysis of the ADBB is not only informative in relation to the construct validity of the full scale but also in relation to how removal of items 4, 6, and 8 in the m-ADBB might have affected the validity of the scale. In addition, in contrast to results derived using CTT, which are sample-specific, results from an IRT approach are, given that IRT assumptions hold, theoretically generalizable beyond the sample studied (Langer et al., 2008; Reeve and Fayers, 2005).

IRT constitutes a set of mathematical models that describe in probabilistic terms the relationship between a person's response to a specific item and his or her level of the latent variable being measured. The latent trait is a hypothetical construct, which is postulated to exist but cannot be directly measured. Instead, the latent trait is indirectly measured using the items comprising the scale (Reeve and Fayers, 2005). Since the ADBB is used as a single dimension, when used for screening purposes, we assess each item's performance in relation to the total score. We use the Two-Parameter Logistic Model (2PLM) to estimate each item's location (or difficulty) and discrimination (or slope) parameters. The functional form is defined as $P_i(x = 1, a_i, b_i) = 1/1 - e - a_i(\theta - b_i)$ with a_i being discrimination and b_i being difficulty. Item location reflects the degree of severity of the latent trait that is needed for a person to have a .50 probability of endorsing the item. Item discrimination (or slope) reflects how well each item identifies cases at different levels of the latent trait with steeper slopes offering better discrimination than less steep slopes. Using location and discrimination parameters, it is possible to examine at which levels of the latent trait specific items discriminate best (Yang and Kao, 2014). Examination of item discrimination and location parameters can be used for evaluating whether some items are redundant or whether more items should be included to best capture the levels of the latent trait of interest. When an item has good discrimination, it becomes less important to have other items measuring the same level of the latent trait. Ideally, a measure should consist of evenly spaced, near vertical item characteristic curves (ICCs) covering the range of the latent trait of interest (Fayers and Machin, 2016).

The use of IRT methods require that the scale fulfill assumptions regarding monotonicity, unidimensionality, absence of differential item functioning (DIF), and local independence. Fulfillment of each of these criteria also provide evidence for the construct validity of the scale. However, IRT models are robust to minor violations of these assumptions (Reeve and Fayers, 2005). Monotonicity implies that the probability of a high item score increases with increasing levels of the latent construct. In relation to the ADBB this means that the probability of showing unusual behavior on one of the items should increase as the level of social withdrawal increases. Unidimensionality implies that items measure a single continuous latent construct. Although previous studies suggest that the ADBB consists of two factors, when conducting ADBB screenings, measures may be considered unidimensional, when there is one dominant factor and relatively minor additional factors (Slocum-Gori et al., 2009). DIF occurs when one or more items perform differently in various subgroups after controlling for overall differences between subgroups on the latent construct (Fayers and Machin, 2016). Presence of DIF suggests that the score on a specific item does not only depend on the latent construct but also on some other characteristic, such as gender or socioeconomic status, which the item was not intended to measure (Langer et al., 2008). On a practical level, presence of DIF may cause problems when different groups are compared using the same measure and/or when the same cutoff is used across different

groups, since DIF may lead to one group erroneously obtaining higher or lower scores hereby potentially affecting the selection rate (percentage of subjects above cutoff) in that group (Hidalgo et al., 2015). Previous studies have shown that social withdrawal in infancy is associated with being a boy (Guedeney et al., 2008) and low gestational age (GA) (Braarud et al., 2013; Guedeney et al., 2012). To ensure that these differences are related to social withdrawal and not DIF, we examine DIF in relation to infant sex and GA. Finally, local independence implies that there is no correlation among items when the latent construct is held constant. This means that although we expect correlations between items in a scale, these correlations should only occur because the items reflect the same latent construct. Local dependence may occur for a number of reasons, e.g. when wording of two items are very similar and raters are not able to differentiate between items, or when items are conditionally dependent on each other (Fayers and Machin, 2016).

In this study, the construct validity of the ADBB is evaluated using IRT in a large community sample with data from public health visitors' routine screenings using the ADBB as part of their daily practice. Evaluating screening instruments when used in a busy real-life setting is an important part of evaluating whether the instrument functions as intended when implemented in practice. In 2015, the ADBB was implemented as a universal screening tool in the public health home visiting program in the City of Copenhagen. ADBB screenings are conducted during routine home visits by public health visitors, i.e. specialized nurses who have completed the "Advanced Nurse Health Visitor Education Program". Health visitors aim at screening infants when they are 2 months, 4-6 months (only in first-time families), and 8-10 months. In addition, a health visitor may choose to conduct further ADBB-screenings in case of extra visits or concern. As a minimum, all infants should be screened with the ADBB at least once during their first year of life (Smith-Nielsen et al., 2018).

2. Methods

2.1. Participants

The study includes data from infants born in the municipality of Copenhagen from March 2014 to July 2019 and their families (N = 55,810). Infants were included in the study, when minimum one ADBB screening was registered between infant age 2-12.9 months (n = 24,752). This is in line with the screening guidelines in the municipality, where all infants should receive a minimum of one screening within the first year postpartum. When two ADBB screenings were registered on an infant with less than 30 days apart, the most recent screening was excluded, since the second screening might be conditional on the first. The ADBB screenings were categorized into three waves corresponding to the age of the infant at the time of the screening: 2-3.9 months, 4-7.9 months, and 8-12.9 months. Each wave includes the infant's first ADBB screening in that age group. An infant may contribute an ADBB screening in more than one wave, as long as there are more than 30 days between the screenings. For a flowchart of the selection process, see Fig. 1.

As a part of the data cleaning process, extreme values of birth weight (< 1.5 kg and > 5.5 kg) and gestational age (< 24 weeks) were excluded.

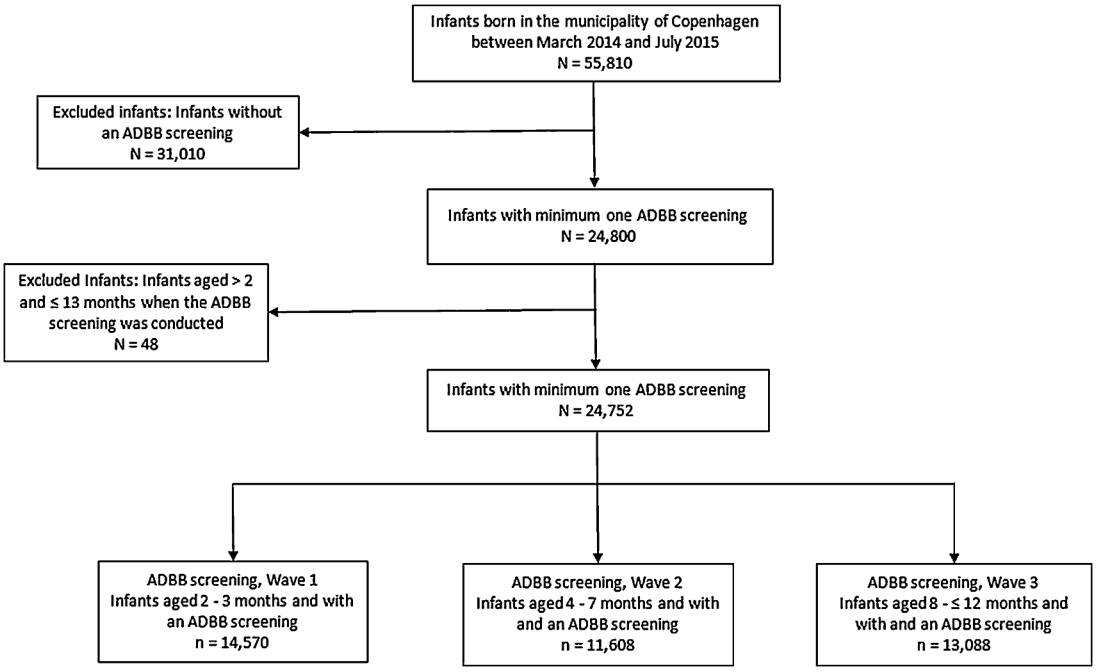

Fig. 1. Flowchart of the study population.

2.2. Procedure

The Danish Health Visitor system is a well-integrated part of the healthcare and welfare system. Data from the home visits are recorded by the public health visitor in the Danish Health Visitor Records. The Danish Civil Registration System (CRS) was used to retrieve information about the study population. Since April 1968, all people living in Denmark are registered by a 10-digit civil person registration (CPR) number in the CRS. Using this number, linkage of individual information from the Danish Health Visitor Records at Novax was possible. From the Danish Health Visitor Records, we obtained information on maternal and paternal age, infant sex, gestational age, date of birth, and ADBB screening, including item scores and date of ADBB screening.

2.3. Alarm Distress Baby Scale

The ADBB (Guedeney and Fermanian, 2001) consists of eight items: (1) Facial expressions, (2) Eye contact, (3) General level of activity, (4) Self-stimulating gestures, (5) Vocalizations, (6) Briskness of response to stimulation, (7) Relationship, and (8) Attraction. Each item is rated on a 5-point scale from 0 (*absolutely normal*) to 4 (*very obvious abnormal behavior*). The total score (range: 0-32) is calculated by summing the scores across the items with higher scores indicating more withdrawal. Studies have agreed on a cutoff of $\geq$ 5 (De Rosa et al., 2010; Facuri-Lopes et al., 2008; Guedeney and Fermanian, 2001; Puura et al., 2010).

The ADBB assessment is conducted during interactions between the infant and a stranger, which in this study was the health nurse. The ADBB can be used in in any context involving social interaction between the observer and the infant, such as pediatric examinations, developmental testing situations or simple face-to-face interactions. When assessing social withdrawal, it is important that the child is in an alert state (Guedeney et al., 2013).

The ADBB is a short screening instrument that was built to be as simple as possible to be implemented and used in clinical practice. It has shown good content validity as well as criterion validity (Guedeney and Fermanian, 2001).

2.4. Training

The ADBB training program of the health visitors were developed in collaboration with Prof. Antoine Guedeney, the developer of the ADBB scale. The health visitors attended a two-day course, where they were given lectures on infant social behavior, an introduction to the ADBB and its use, and a thorough introduction to scoring criteria. The health visitors also practiced infant observation using the ADBB scale with videotaped examples. Following the course, the health visitors independently rated 11 videos. Their ratings were discussed and compared with expert ratings during three two-hour group supervision sessions. Before the health visitors began to use the ADBB scale in their practice, they had to demonstrate acceptable agreement ($\geq$ 75%) on a reliability set consisting of four videos. To rate a video correctly, the rater's total score should fall into one of the following categories, also listed in the ADBB manual (Guedeney, 2015): total score = 0–4 ('No concern'); total score = 5–10 ('Some concern'); total score $\geq$ 11 ('Significant concern'). To keep up the obtained reliability and avoid coders' drift, health visitors attended a two-hour group supervision session every third month on their use of the ADBB scale. At the group supervisions, the scoring of videotaped ADBB-screenings conducted either by the health visitors or the supervisor were discussed. All trainers had demonstrated acceptable agreement by rating six out of seven videos correctly in relation to the three categories (corresponding to Cohen's $\kappa \geq .78$ and $\geq$ 85% agreement) and by showing correct item ratings (reference score +/- 1) on 90% of the items. Certification videos were rated by Prof. A. Guedeney.

2.5. Statistical analyses

All analyses were conducted in R v. 3.6.3. The scores on each item were transformed to a binary score: Scores of zero where kept indicating that the infant had no unusual behavior on this item, and scores of one to four were transformed to a score of one indicating that the infant showed some degree of unusual behavior on that item. The total score was calculated as the sum of the binary items. This was done due to the very low prevalence of scores above 0 in our sample (76.4% of the sample had a total score of 0 across the three waves).

Firstly, the assumptions regarding monotonicity, unidimensionality, local independence, and no differential item functioning (DIF) were tested. Monotonicity was examined by inspecting the average item score in relation to the score obtained by summing all the items except the item in question. The R package ltm (Rizopoulos, 2006) was used to test for unidimensionality. In this package, unidimensionality is tested using modified parallel analysis, a method developed by Drasgow and Lissak (1983) for examining the latent dimensionality of dichotomously scored items. Local dependence was examined using correlations between items while controlling for the rest score, i.e. the total score minus the two items. Pearson correlations (r_p) exceeding 0.2 are suggestive of local dependence (Fayers and Machin, 2016). DIF was examined by fitting logistic regression models of each item on the total score and the exogenous variable. DIF was examined in relation to infant sex and gestational age. DIF was considered present when the effect reached statistical significance, and log-odds ratios (log-OR) were outside the +/- 0.64 interval (Scott et al., 2010).

Secondly, the R package ltm (Rizopoulos, 2006) was used to evaluate item fit to a 2PL IRT model and estimate item parameters. For evaluating item fit, a Monte Carlo procedure were used to approximate the distribution of the item-fit statistic under the null hypothesis. The following categories are used to describe item discrimination parameters: Very low ($0.01 < \alpha_i < 0.34$), low ($0.35 < \alpha_i < 0.64$), moderate ($0.65 < \alpha_i < 1.34$), high ($1.35 < \alpha_i < 1.69$), and very high ($\alpha_i > 1.70$) (Baker and Kim, 2017). If an item does not fit the model, it may be due to e.g. poor item quality or poor construct validity.

Given that several statistical tests were performed and the large sample size, results were evaluated at the 1% significance level.

2.6. Ethical considerations

The Danish Data Protection Agency, through the notification system at Copenhagen University, approved the usage of the data obtained in The Copenhagen Infant Mental Health Project (journal number: 514-0026/18-2000). All data have be pseudo anonymized before statistical analyses and publications according to Danish law. By using unique IDs anonymity is secured for all retrieved data. The sheet including personal identifiable information is securely stored separately from data and only authorized personnel from the project have access to this sheet and to the data via a secure interface. Investigators only have access to data sets cleaned of personal identifiable information. By Danish law, no informed consent is required for a registry-based study.

3. Results

3.1. Descriptive statistics

Between 11,608 and 14,570 infants were screened at each wave. Fewest infants were screened at wave two, which is the wave where only first-time parents receive routine visits. Descriptive statistics were calculated for Wave 1, 2 and 3 and are presented separately for each Wave in Table 1. Item 5 was the item where most infants showed some degree of unusual behavior, while item 4 and item 6 were the items where fewest infants showed some degree of unusual behavior.

3.2. Assessment of construct validity

First, we tested unidimensionality for the full scale at each of the three waves. Unidimensionality was not rejected for the full scale at wave two ($p = .020$) and three ($p = .030$), but unidimensionality was rejected for the full scale at wave one ($p = .0099$).

Then we examined the assumption regarding local independence. Across the three waves, item 7 and item 8 showed strong local dependence ($r_p = .510$). In addition, item 7 showed local dependence with item 2 ($r_p = .298$), and item 8 showed local dependence with item 2 ($r_p = .232$). See Tables 1–3 in the supplementary material for partial correlation matrices at each wave.

Violation of the assumption of local independence may affect the estimation of item parameters, especially overestimation of discrimination parameters leading to very steep slopes ($\alpha > 4$) (Edelen and Reeve, 2007; Nguyen et al., 2014). Based on this some authors suggest that items showing local dependence above 0.25 should be removed (Rose et al., 2008). Other studies suggest that removal of misfitting items only improves results in cases of severe multidimensionality and a large proportion of misfitting items, and results deteriorate otherwise (Crişan et al., 2017). When estimating item discrimination parameters, we payed special attention to the issue related to inflated discrimination parameters ($\alpha > 4$) for items 2, 7, and 8.

Finally, we tested the presence of DIF in relation to infant sex and gestational age. In waves 1 and 2 a degree of DIF was suggested for infant's sex (reference category = boy). The two items for wave 1 were item 2 (log-OR = -0.265, p = .003) and item 5 (log-OR = 0.276, p < .001). For wave 2 they were item 8 (log-OR = -0.385, p = .001) and item 5 (log-OR = 0.255, p < .001). There are no statistically significant effects at wave three (see Table 4 in the supplementary materials for log-OR and all DIF tests).

Table 1
Participant information and ADBB scores across the three waves.

	Wave 1 $n = 14{,}570$	Wave 2 $n = 11{,}608$	Wave 3 $n = 13{,}088$
Sex, girl, % (*n*)	49 (7,099)	49 (5,375)	49 (6,407)
Gestational age, weeks, *M* (SD)	39.8 (1.6)	39.7 (2.1)	39.7(2.0)
Missing, *n*	254	283	380
Birth weight, grams, *M* (SD)	3473 (526)	3417 (538)	3458 (537)
Missing, *n*	172	294	356
Maternal age, years, *M* (SD)	32.2 (4.6)	31.7 (4.7)	32.1 (4.7)
Missing, *n*	37	13	10
Paternal age, years, *M* (SD)	34.4 (5.8)	33.7 (5.8)	34.3 (5.7)
Missing, *n*	416	359	332
Infant age, months, *M* (SD)	2.4 (0.5)	5.0 (1.2)	8.6 (0.6)
Prevalence of ADBB scores			
1. Facial expressions ≠ 0, % (*n*)	6.6 (963)	5.4 (628)	4.2 (546)
2. Eye contact ≠ 0, % (*n*)	7.6 (1111)	2.6 (304)	1.6 (210)
3. General activity level ≠ 0, % (*n*)	4.0 (588)	3.7 (429)	2.3 (300)
4. Self-stimulating gestures ≠ 0, % (*n*)	0.7 (100)	0.9 (109)	0.7 (94)
5. Vocalizations ≠ 0, % (*n*)	15.8 (2295)	16.9 (1964)	13.1 (1710)
6. Briskness of response to stimulation ≠ 0, % (*n*)	1.2 (179)	0.8 (93)	0.7 (88)
7. Relationship ≠ 0, % (*n*)	9.6 (1396)	5.5 (633)	4.4 (582)
8. Attraction ≠ 0, % (*n*)	10.5 (1525)	6.1 (707)	4.2 (554)

Note. ADBB = Alarm Distress Baby Scale. ≠ 0 = the percentage of infants with slightly to very unusual behavior.

Table 2
Item parameters and item-fit at wave 1.

		Location parameter (*b*)			Discrimination parameter (*a*)			Item-fit
		Value	Std. Err.	Z-value	Value	Std. Err.	Z-value	*p*-value
1	Facial expressions	2.09	0.05	43.95	1.76	0.07	26.64	.089
2	Eye contact	1.56	0.02	69.29	3.51	0.15	23.04	.614
3	General level of activity	3.29	0.14	23.19	1.11	0.06	18.12	.020
4	Self-stimulating gestures	4.42	0.36	12.21	1.29	0.14	9.49	.624
5	Vocalizations	1.71	0.04	38.43	1.22	0.04	28.95	.614
6	Briskness of response to stimulation	3.19	0.13	23.74	1.78	0.12	15.03	.901
7	Relationship	1.35	0.01	185.37	12.14	3.85	3.16	.010
8	Attraction	1.33	0.01	95.30	4.56	0.22	20.86	.168

Table 3
Item parameters and item-fit at wave 2.

		Location parameter (*b*)			Discrimination parameter (*a*)			Item-fit
		Value	Std. Err.	Z-value	Value	Std. Err.	Z-value	*p*-value
1	Facial expressions	2.22	0.05	41.06	1.90	0.08	22.53	.040
2	Eye contact	2.25	0.04	51.43	3.13	0.18	17.20	.891
3	General level of activity	3.00	0.12	25.32	1.35	0.08	17.53	.545
4	Self-stimulating gestures	4.45	0.37	11.93	1.20	0.13	9.19	.842
5	Vocalizations	1.59	0.05	34.74	1.31	0.05	24.62	.188
6	Briskness of response to stimulation	3.55	0.19	18.31	1.76	0.16	11.35	.020
7	Relationship	1.76	0.03	64.45	4.47	0.26	17.19	.584
8	Attraction	1.72	0.03	63.22	4.05	0.22	18.50	.713

Table 4
Item parameters and item-fit at wave 3.

		Location parameter (*b*)			Discrimination parameter (*a*)			Item-fit
		Value	Std. Err.	Z-value	Value	Std. Err.	Z-value	*p*-value
1	Facial expressions	2.30	0.16	14.26	2.11	0.09	22.81	.010
2	Eye contact	2.50	0.28	9.02	3.07	0.20	15.51	.980
3	General level of activity	3.05	0.25	12.05	1.61	0.09	17.32	.396
4	Self-stimulating gestures	4.31	0.60	7.23	1.35	0.14	9.40	.317
5	Vocalizations	1.85	0.11	17.08	1.34	0.05	24.52	.059
6	Briskness of response to stimulation	3.60	0.46	7.90	1.82	0.16	11.37	.891
7	Relationship	1.83	0.22	8.19	5.98	0.37	16.02	.723
8	Attraction	1.88	0.19	9.70	4.95	0.27	18.62	.782

3.3. Item parameters and item fit

The results from the 2PLM with all items included showed that item 7 and item 8 had very steep slopes across the three waves ($a_i \geq 4.05$). To assess the influence of the local dependence between this item-pair on the model, we compared item parameters and ICCs for (a) the 2PLM with all items, (b) a 2PLM without item 7, and (c) a 2PLM without item 8 (Toland, 2014).

Removal of either item 7 or item 8 from the model did not lead to large changes in location parameters, but as expected it did lead to large reductions in the discrimination parameters for item 7 and item 8, respectively. Yet, the discrimination parameters for these items remained high ($as \geq 2.59$). In addition, examination of the ICCs showed that the items had similar shapes across the three waves with and without either item 7 or item 8 in the model. Since location parameters were not heavily influenced and ICCs remained similar, we decided to keep all items in the model. When interpreting the results, it should be kept in mind that discrimination parameters for items 7 and 8 might be overestimated.

Item parameters and fit statistics are displayed in Tables 2-4 (for ICCs at each wave, see supplementary material, Fig.s 1-3). The majority of the items fit the 2PLM (all $ps > .01$). But in wave one, the shape of the ICC for item 7 does not fit the 2PLM ($p = .0099$), and in wave three, the shape for item 1 does not fit the 2PLM ($p = .0099$) (see Fig. 1 and 3 in the supplementary material, respectively, for the ICCs). Since ICCs showed that items had a similar pattern across the three waves, and there were no pattern in the items not fitting the model, we consider the marginal significant *p*-values were a result of randomness and not as problematic deviations.

The items displayed similar patterns in terms of location and discrimination parameters across the three waves. All location parameters were above one, indicating that the items generally discriminate best above average levels of social withdrawal. The items seem to be clustered in two groups, where items 1, 2, 5, 7, and 8 ($1.33 \leq b_i \leq 2.25$) generally discriminate better at lower levels of the latent trait compared to items 3, 4, and 6 ($3.00 \leq b_i \leq 4.45$). The discrimination parameters indicate that the items show moderate to good discriminatory abilities ($a_i \geq 1.11$). Items 1, 2, 7, and 8 were among the most discriminating items across the three waves ($a_i \geq 1.76$). Items 3, 4, and 5 were among the least discriminating items across the three waves ($1.11 \leq a_i \leq 1.61$).

4. Discussion

Using IRT methods, the aim of this study was to examine the construct validity of the Alarm Distress Baby Scale (ADBB) when used in a busy real-life primary care setting. Overall, the items fulfilled IRT assumptions regarding unidimensionality, monotonicity, and no statistical and clinical significant DIF. But fulfilment of local independence for all items was not achieved. In addition, results showed that the items discriminate well and discriminate better above average levels of social withdrawal.

Analyses of unidimensionality suggested that the eight items on the ADBB were sufficiently unidimensional to be used as a single dimension as the tests for wave two and three were insignificant, and for wave one it just reached the threshold of the significance level. However, it is possible that the ADBB in fact contains two factors, which has been suggested by previous studies (Guedeney et al., 2013). Based on the current data, one of these factors seems sufficiently dominant to not reject the assumption of unidimensionality. Thus, it seems reasonable to use the total score of the ADBB during screenings. Further, the items show a monotonic relationship to the latent trait, i.e. the probability of showing problematic behavior on one of the items increases as the level of social withdrawal increases. Finally none of the items showed clinical and statistically significant DIF, suggesting that there are no major problems with DIF, indicating that the items function similarly across infant's sex and gestational age. Since the ADBB is used as a universal screening tool, we consider this an important psychometric property. When the same cut-off scores are used across different groups of individuals, it is of high importance that items does not show DIF, since DIF might affect the selection rates for a specific group (Hidalgo et al., 2015). The practical consequences of the items showing statistically significant DIF in relation to infant's sex however should be considered, when the cut-off is validated in a Danish setting.

In contrast, the assumption regarding local independence was not met. There was strong local dependence between item 7 (Relationship) and item 8 (Attraction) across the three waves. In addition, items 7 and 8 also showed local dependence with item 2 (Eye contact). Local dependence may occur for different reasons, e.g. when items are similar making it difficult for respondents to differentiate between the items, or when items have shared variance that is not due to the latent factor (Edelen and Reeve, 2007; Toland, 2014). In the following, we focus on local dependence between items 7 and 8, since this item-pair showed the strongest local dependence. The item content of these items are defined as the infant's ability to engage in a relationship with the observer (item 7) and the infant's ability to attract and maintain the observer's attention (item 8). As such, ratings of items 7 and 8 occur at a more general level, in contrast to ratings of the other items which assess specific types of behavior, such as facial expressions or vocalizations. In addition, during the training and ongoing supervision of the health visitors, difficulties related to differentiating between items 7 and 8 were commonly raised as an issue. The strong local dependence between items 7 and 8 suggest that when used by public health visitors in a busy real-life setting, the construct validity of the ADBB could be improved by removing one of these items as suggested in the m-ADBB.

Parameter estimates from the 2PLMs show similarities in terms of location and discrimination across the three waves, which suggest that the ADBB items function similarly when assessing social withdrawal in infants aged 2-13 months. As such, infant age does not affect the validity of the ADBB items.

The items' location parameters are useful for examining whether the items capture the full range of the latent trait, or the range of interest in relation to the specific instrument. The relative item location parameters are considered 'sample free', meaning that the ordering of the items' locations along the latent trait should remain stable across samples (Fayers and Machin, 2016). The ADBB items discriminate best above average levels of social withdrawal ($b_i \geq 1.46$). Since the ADBB is a screening instrument, the purpose of the items is not to show good discrimination at average or below average levels of social withdrawal, but to discriminate between infants showing average and above average levels of social withdrawal. Based on this, the spread of location parameters above average levels of social withdrawal is considered optimal in a screening context (Baker and Kim, 2017). Although, it could be argued that items showing good discrimination at very high levels of social withdrawal might be of limited relevance in a screening context. Therefore item 4 (Self-stimulating gestures) and item 6 (Briskness of response to stimulation) might be of limited use in a screening context due to their very large location parameters ($b_i \geq 3.20$). In the m-ADBB, Matthey and colleagues (2013) have removed these two items but for other reasons, i.e. low interrater reliability on item 4 (Self-stimulating gestures) and high correlations between item 6 (Briskness of response to stimulation) and item 3 (General level of activity). Based on results from the present study, it could be argued that these two items do not provide enough information at the levels of social withdrawal of interest for screening purposes. On the other hand, it could also be argued that it is useful for clinicians to be able to distinguish between infants with high and very high levels of social withdrawal already during the screening. Further, items 4 and 6 could also be important risk markers for specific disorders in infancy. Previous studies have shown that self-stimulating behaviors, such as stereotypies, are common for children with autism and other developmental disabilities (Barbaro and Dissanayake, 2009), and a recent study using the ADBB showed that infants with Prader-Willi syndrome show delayed responses to stimulation (Tauber et al., 2017). However, the predictive and criterion-related validity of items 4 and 6 in relation to detecting specific disorders should be examined in future studies.

The discrimination parameter informs on the item's ability to separate individuals below and above the item's location along the latent trait. Items with higher discrimination parameters are better at differentiating between individuals below and above the item's location parameter (Fayers and Machin, 2016). The results showed that all items showed moderate to good discriminatory abilities ($a_i \geq 1.11$). Item 5 (Vocalizations) was among the least discriminating item across the three waves ($a_i \leq 1.35$). Although the discriminatory ability of item 5 is acceptable, the ICCs show that item 5 discriminates within the same range of social withdrawal as items 7 (Relationship) and 8 (Attraction). Based on this, it could be argued that item 5 does not provide sufficiently new information to be retained in the scale. On the other hand, it could also be argued that it is important to retain item 5, since the content of this item (amount of vocalizations) is not covered by any other items.

In addition, the ICCs show that the shapes of items 7 and 8 are almost identical across the three waves. This suggest that items 7 and

8 are equally good at differentiating between infants with the same level of social withdrawal. Based on this, it could be argued that one of these items should be removed, since the items do not provide sufficient unique information relative to each other. Taken together with the strong local dependence between these items, our results provide another line of empirical support for Matthey and colleagues' (2013) decision to remove item 8 in the m-ADBB. In practice, the almost identical ICCs for items 7 and 8 means that when scores from the ADBB items are summed to a total score, changes around the level of social withdrawal measured by items 7 and 8 are weighted double. How this affects the criterion-related validity of the scale should be examined in future studies.

Declaration of Competing Interest

The authors declare that there is a potential conflict of interest, since the Centre for Early Intervention and Family Studies, where five of the authors are employed, offers training in the use of the ADBB scale as part of the continuing education program at the University of Copenhagen.

Limitations and strengths

Limitations need to be taken into account when interpreting results from the present study. First, the low prevalence of social withdrawal and especially the low prevalence of scores above zero on certain items, i.e. item 4 (Self-stimulating gestures) and item 6 (Response time), may limit the reliability of the estimated item parameters (Fayers and Machin, 2016). Second, there was detected local dependence between items 2 (Eye contact), 7 (Relationship), and 8 (Attraction), and especially between items 7 and 8. The items were kept in the model, since ICCs were similar with and without either item 7 or 8 in the model. There are other possible ways of dealing with local dependence which might have yielded different results, e.g. using more complicated IRT models accounting for the local dependence between item-pairs (Choi and Asilkalkan, 2019). Third, while we did chose a more conservative significance level of 1%, the large sample size and number of statistical tests may have influenced the results. As sample size increases so does the statistical power, making it more likely to reject the null-hypothesis as long as the population effect size is not exactly zero (Faber and Fonseca, 2014; Sullivan and Feinn, 2012), and we may not have been conservative enough. Throughout our tests, we have relied on interpreting effect sizes that are not sensitive to sample size except for when testing for unidimensionality. However, as already mentioned, the tests were insignificant or just reached the threshold of significance. Taken together with the large sample size and the multiple number of tests, this indicates that lack of unidimensionality is at worst a very minor issue and most likely not an issue at all.

We also believe the study has several strengths. First, this is the first study to examine the construct validity of the ADBB in a real-world setting. In research, there has been a long discussion about the knowledge-practice gap, and research-based methods have been developed in an attempt at bridging this gap (Westerlund et al., 2019). The validity of these methods are typically examined during development in research settings, and then transferred to practice. But when the methods meet the demands of real-life, their validity might be compromised. Results from this study sheds light on the psychometric strengths and weaknesses of the ADBB when used in a real-world setting by non-expert raters. Second, this is the first study to examine DIF in relation to the items on the ADBB. It is important to examine DIF, when screening instruments use the same cut-off scores for different groups of individuals, since the presence of DIF might erroneously affect selection rates in specific groups (Hidalgo et al., 2015). Third, this is also the first study to use IRT to examine the psychometric properties of the ADBB. Compared to CTT, IRT allows a more in-depth analysis of how the items function. We believe our results are useful in relation to deciding in which situations to use the full and the m-ADBB. Finally, this is the first study to examine the validity of the ADBB across three large age groups, hereby providing information on the extent to which the ADBB is equally valid at different infant ages.

Implications for practice

The ADBB shows several psychometric strengths when used by public health visitors in primary care, and the items generally show good discriminatory abilities at the levels of social withdrawal of interest for screening purposes. The ADBB also shows some psychometric weaknesses. Based on these weaknesses, we argue that our results may be viewed as support for the m-ADBB in terms of removal of item 8 (Attraction), since this item does not provide sufficient unique information in relation to item 7 (Relationship), and removal of items 4 (Self-stimulating gestures) and 6 (Briskness of response to stimulation), since these items discriminate best at very high levels of social withdrawal, which might be of limited interest for screening purposes. However, it is important for clinicians to be aware that items 4 and 6 might be informative when assessing infants with specific disorders, such as Autism Spectrum Disorder or Prader-Willi syndrome. Therefore, the m-ADBB could be used as a first-line screening instrument, and the full ADBB could be used when assessing social withdrawal in specific at-risk samples. Yet, before implementing the m-ADBB as a screening tool in primary care, it is important that future studies compare the criterion-related validity of the full and the m-ADBB, since an instrument's ability to correctly detect infants at-risk is the most important psychometric property in a screening context.

Funding

The project is funded by a grant from the charitable foundation Tryg Foundation (Grant ID no 107616).

Acknowledgments

The authors thank the health visitors and The Children and Youth Administration (Børne- og Ungdomsforvaltningen) in the City of Copenhagen for a fruitful collaboration. The authors also acknowledge the valuable contribution of Rie Krondorf von Wowern-Davies to the translation of the ADBB-manual into Danish and assistance in the development of the training seminar. Finally, the authors thank the psychologists who have conducted the training and ongoing supervision of the health visitors throughout the project.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijnsa.2021.100038.

References

Assumpção Jr., F.B., Kuczynski, E., Rego, M.G.da S., Rocca, C.C.de A, 2002. Escala de avaliação da reação de retração no bebê: Um estudo de validade. Arq. Neuropsiquiatr. 60 (1), 56–60. https://doi.org/10.1590/S0004-282X2002000100011.

Bagner, D.M., Rodríguez, G.M., Blake, C.A., Linares, D., Carter, A.S., 2012. Assessment of behavioral and emotional problems in infancy: a systematic review. Clin. Child Fam. Psychol. Rev. 15 (2), 113–128. https://doi.org/10.1007/s10567-012-0110-2.

Baker, F.B., Kim, S.H., 2017. The Basics of Item Response Theory Using R. Springer.

Barbaro, J., Dissanayake, C., 2009. Autism spectrum disorders in infancy and toddlerhood: a review of the evidence on early signs, early identification tools, and early diagnosis. J Develop. Behav. Pediatrics 30 (5), 447–459. https://doi.org/10.1097/DBP.0b013e3181ba0f9f.

Braarud, H.C., Slinning, K., Moe, V., Smith, L., Vannebo, U.T., Guedeney, A., Heimann, M., 2013. Relation between social withdrawal symptoms in full-term and premature infants and depressive symtoms in mothers: A longitudinal study. Infant Mental Health J. 34 (6), 532–541. https://doi.org/10.1002/imhj.21414.

Brazelton, B.T., Kolowski, B., Main, M., 1974. The origins of reciprocity. In M. Lewis and L. D. Rosenblum (Eds.). The Effect of the Infant on its Caregiver. Wiley-Interscience, pp. 137–154.

Briggs-Gowan, M.J., Carter, A.S., Bosson-Heenan, J., Guyer, A.E., Horwitz, S.M., 2006. Are infant-toddler social-emotional and behavioral problems transient? J. Am. Acad. Child and Adolesc. Psychiatry 45 (7), 849–858. https://doi.org/10.1097/01.chi.0000220849.48650.59.

Choi, Y.-J., Asilkalkan, A., 2019. R Packages for item response theory analysis: descriptions and features. Measurement 17 (3), 168–175. https://doi.org/10.1080/15366367.2019.1586404.

Crişan, D.R., Tendeiro, J.N., Meijer, R.R., 2017. Investigating the practical consequences of model misfit in unidimensional IRT models. Appl. Psychol. Measur. 41 (6), 439–455. https://doi.org/10.1177/0146621617695522.

De Rosa, E., Curró, V., Wendland, J., Maulucci, S., Maulucci, M.L., De Giovanni, L., 2010. Propriétés psychométriques de l'échelle Alarme Détresse Bébé (ADBB) appliquée à 81 enfants italiens. Devenir 22 (3), 209. https://doi.org/10.3917/dev.103.0209.

Drasgow, F., Lissak, R.I., 1983. Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. J. Appl. Psychol. 68 (3), 363–373. https://doi.org/10.1037/0021-9010.68.3.363.

Edelen, M.O., Reeve, B.B., 2007. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual. Life Res. 16 (SUPPL. 1), 5–18. https://doi.org/10.1007/s11136-007-9198-0.

Faber, J., Fonesca, L.M., 2014. How sample size influence research outcomes. Dental Press J. Orthodontics 19 (4), 27–29.

Facuri-Lopes, S.C., Ricas, J., Mancini, M.C., 2008. Evaluation of the psychometrics properties of the alarm distress baby scale among 122 Brazilian children. Infant Mental Health J. 29 (2), 153–173. https://doi.org/10.1002/imhj.20169.

Fayers, P.M., Machin, D., Fayers, P.M., Machin, D., 2016. Item response theory and differential item functioning. Quality of Life, 3rd Ed. John Wiley and Sons, pp. 161–188. https://doi.org/10.1002/9780470024522.ch7.

Field, T.M., 1981. Infant gaze aversion and heart rate during face-to-face interactions. Infant Behav. Devel. 4 (1), 307–315. https://doi.org/10.1016/S0163-6383(81)80032-X.

Groh, A.M., Roisman, G.I., van IJzendoorn, M.H., Bakermans-Kranenburg, M.J., Fearon, R.P., 2012. The significance of insecure and disorganized attachment for children's internalizing symptoms: a meta-analytic study. Child Dev. 83 (2), 591–610. https://doi.org/10.1111/j.1467-8624.2011.01711.x.

Guedeney, A., 2015. Alarm Distress Baby Scale: A recent Manual for Use. Unpublished manual.

Guedeney, A., Doukhan, S., Forhan, A., Heude, B., Peyre, H., 2017. To which extent social withdrawal at the age of 1 year is associated with IQ at 5–6 years old? Results of the EDEN mother–child cohort. European Child and Adolescent Psychiatry 26 (11), 1343–1350. https://doi.org/10.1007/s00787-017-0988-9.

Guedeney, A., Fermanian, J., 2001. A validity and reliability study of assessment and screening for sustained withdrawal reaction in infancy: The Alarm Distress Baby Scale. Infant Mental Health J. 22 (5), 559–575.

Guedeney, A., Forhan, A., Larroque, B., de Agostini, M., Pingault, J.-B., Heude, B., 2016. Social withdrawal behaviour at one year of age is associated with delays in reaching language milestones in the EDEN mother-child cohort study. PLoS One 11 (7), e0158426. https://doi.org/10.1371/journal.pone.0158426.

Guedeney, A., Foucault, C., Bougen, E., Larroque, B., Mentré, F., 2008. Screening for risk factors of relational withdrawal behaviour in infants aged 14–18 months. Eur. Psychiatry 23 (2), 150–155. https://doi.org/10.1016/j.eurpsy.2007.07.008.

Guedeney, A., Marchand-Martin, L., Cote, S.J., Larroque, B., 2012. Perinatal risk factors and social withdrawal behaviour. Eur. Child and Adoles. Psychiatry 21 (4), 185–191. https://doi.org/10.1007/s00787-012-0250-4.

Guedeney, A., Matthey, S., Puura, K., 2013. Social withdrawal behavior in infancy: a history of the concept and a review of published studies using the alarm distress baby scale. Infant Mental Health J. 34 (6), 516–531. https://doi.org/10.1002/imhj.21412.

Guedeney, A., Pingault, J.-B., Thorr, A., Larroque, B., 2014. Social withdrawal at 1 year is associated with emotional and behavioural problems at 3 and 5 years: the Eden mother-child cohort study. Eur. Child Adolesc. Psychiatry 23 (12), 1181–1188. https://doi.org/10.1007/s00787-013-0513-8.

Guevara, J.P., Gerdes, M., Localio, R., Huang, Y.V., Pinto-Martin, J., Minkovitz, C.S., Hsu, D., Kyriakou, L., Baglivo, S., Kavanagh, J., Pati, S., 2013. Effectiveness of developmental screening in an urban setting. Pediatrics 131 (1), 30–37. https://doi.org/10.1542/peds.2012-0765.

Hidalgo, M.D., Galindo-Garre, F., Gómez-Benito, J., 2015. Differential item functioning and cut-off scores: implications for test score interpretation*. UB J. Psychol. 45 (1), 55–69. www.redalyc.org/articulo.oa?id=970/97041174004.

Kyriazos, T.A., 2018. Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. Psychology 09 (08), 2207–2230. https://doi.org/10.4236/psych.2018.98126.

Langer, M.M., Hill, C.D., Thissen, D., Burwinkle, T.M., Varni, J.W., DeWalt, D.A., 2008. Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures. J. Clin. Epidemiol. 61 (3), 268–276. https://doi.org/10.1016/j.jclinepi.2007.05.002.

Matthey, S., Črnčec, R., Hales, A., Guedeney, A., 2013. A description of the modified alarm distress baby scale (m-ADBB): an instrument to assess for infant social withdrawal. Infant Mental Health J. 34 (6), 602–609. https://doi.org/10.1002/imhj.21407.

Matthey, S., Guedeney, A., Starakis, N., Barnett, B., 2005. Assessing the social behavior of infants: use of the ADBB Scale and relationship to mother's mood. Infant Mental Health J. 26 (5), 442–458. https://doi.org/10.1002/imhj.20061.

Milne, L., Greenway, P., Guedeney, A., Larroque, B., 2009. Long term developmental impact of social withdrawal in infants. Infant Behav. Develop. 32 (2), 159–166. https://doi.org/10.1016/j.infbeh.2008.12.006.

Nguyen, T.H., Han, H.R., Kim, M.T., Chan, K.S., 2014. An introduction to item response theory for patient-reported outcome measurement. Patient 7 (1), 23–35. https://doi.org/10.1007/s40271-013-0041-0.

Pérez-Martínez, C., Grollemund, B., Gavelle, P., Frías-Navarro, M.D., Alfaiate, T., Mullaert, J., Guedeney, A., 2020. Comparing the Alarm Distress Baby Scale (ADBB) and the Modified version of the ADBB Scale (m-ADBB) to assess social withdrawal behavior in infants with Cleft Lip and Palate. *Submitted for Publication*.

Puura, K., Mäntymaa, M., Luoma, I., Kaukonen, P., Guedeney, A., Salmelin, R., Tamminen, T., 2010. Infants' social withdrawal symptoms assessed with a direct infant observation method in primary health care. Infant Behav. Develop. 33 (4), 579–588. https://doi.org/10.1016/j.infbeh.2010.07.009.

Reeve, B.B., Fayers, P.M, 2005. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers, P.M., Hays, R. (Eds.), Assessing Quality of Life in Clinical Trials: Methods and Practice, 2nd ed. Oxford University Press, pp. 55–73.

Rizopoulos, D., 2006. ltm: An R package for latent variable modeling and item response theory analyses. J. Statistic. Software 17 (5), 1–25.

Rose, M., Bjorner, J.B., Becker, J., Fries, J.F., Ware, J.E., 2008. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J. Clin. Epidemiol. 61 (1), 17–33. https://doi.org/10.1016/j.jclinepi.2006.06.025.

Sand, N., 2005. Pediatricians' Reported Practices Regarding Developmental Screening: Do Guidelines Work? Do They Help? Pediatrics 116 (1), 174–179. https://doi.org/10.1542/peds.2004-1809.

Scott, N.W., Fayers, P.M., Aaronson, N.K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M.A., Sprangers, M.A., 2010. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual. Life Outcomes 8 (1), 81. https://doi.org/10.1186/1477-7525-8-81.

Seifer, R., Sameroff, A., Dickstein, S., Schiller, M., Hayden, L.C., 2004. Your own children are special: clues to the sources of reporting bias in temperament assessments. Infant Behav. Develop. 27 (3), 323–341. https://doi.org/10.1016/j.infbeh.2003.12.005.

Shonkoff, J.P., Phillips, D., 2000. From neurons to neighborhoods the science of early child development. National Academy Press.

Slocum-Gori, S.L., Zumbo, B.D., Michalos, A.C., Diener, E., 2009. A note on the dimensionality of quality of life scales: an illustration with the satisfaction with life scale (SWLS). Social Indicat. Res. 92 (3), 489–496. https://doi.org/10.1007/s11205-008-9303-y.

Smith-Nielsen, J., Lønfeldt, N., Guedeney, A., Væver, M.S., 2018. Implementation of the Alarm Distress Baby Scale as a universal screening instrument in primary care: feasibility, acceptability, and predictors of professionals' adherence to guidelines. Int. J. Nurs. Stud. 79, 104–113. https://doi.org/10.1016/j.ijnurstu.2017.11.005.

Smith-Nielsen, J., Lange, T., Wendelboe, K.I., Wowern, R.K., Væver, M.S., 2019. Associations between maternal postpartum depression, infant social behavior with a stranger, and infant cognitive development. Infancy 24 (4), 663–670. https://doi.org/10.1111/infa.12287.

Stifter, C.A., Willoughby, M.T., Towe-Goodman, N., 2008. Agree or agree to disagree? Assessing the convergence between parents and observers on infant temperament. Infant. Child Develop. 17 (4), 407–426. https://doi.org/10.1002/icd.584.

Sullivan, G.M., Feinn, R., 2012. Using effect size – or why the P value is not enough. J. Graduat. Med. Educ. 4 (3), 279.

Tauber, M., Boulanouar, K., Diene, G., Çabal-Berthoumieu, S., Ehlinger, V., Fichaux-Bourin, P., Molinas, C., Faye, S., Valette, M., Pourrinet, J., Cessans, C., Viaux-Sauvelon, S., Bascoul, C., Guedeney, A., Delhanty, P., Geenen, V., Martens, H., Muscatelli, F., Cohen, D., Salles, J.P., 2017. The use of oxytocin to improve feeding and social skills in infants with prader-willi syndrome. Pediatrics 139 (2). https://doi.org/10.1542/peds.2016-2976.

Tavakol, M., Dennick, R., 2011. Making sense of Cronbach's alpha. Int. J. Med. Educ. 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd.

Toland, M.D., 2014. Practical guide to conducting an item response theory analysis. J. Early Adolesc. 34 (1) https://doi.org/10.1177/0272431613511332.

Ulak, M., Ranjitkar, S., Shrestha, M., Braarud, H.C., Chandyo, R.K., Shrestha, L., Guedeney, A., Strand, T.A., Kvestad, I., 2020. The feasibility of the full and modified versions of the alarm distress baby scale (ADBB) and the prevalence of social withdrawal in infants in Nepal. Front. Psychol. 11 (August), 1–10. https://doi.org/10.3389/fpsyg.2020.02025.

Westerlund, A., Nilsen, P., Sundberg, L., 2019. Implementation of implementation science knowledge: the research-practice gap paradox. Worldviews Evid.-Based Nurs. 16 (5), 332–334. https://doi.org/10.1111/wvn.12403.

Yang, F.M., Kao, S.T., 2014. Item response theory for measurement validity. Shanghai Arch. Psychiatry 26 (3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.

Zhou, F., Huang, P., Wei, X., Guo, Y., Lu, J., Feng, L., Lu, M., Liu, X., Tu, S., Deprez, A., Guedeney, A., Shen, S., Qiu, X., 2020. Prevalence and Characteristics of Social Withdrawal in Chinese Young Children: A Pilot Study. *Submitted for Publication*.