The Director Task Fails to Differentiate Young Adult Theory of Mind Abilities: An IRT Analysis

Mikhail Sokolov (MishaSokolov@cmail.carleton.ca)
John Logan (JohnLogan@cunet.carleton.ca)
Department of Psychology, Carleton University
1125 Colonel By Drive Ottawa, ON K1S5B6 Canada

Abstract

The goal of the present study was to demonstrate the potential application of Item Response Theory (IRT) outside its traditional use in assessing questionnaires by applying it to data from behavioural task. We did this by validating a perspective taking task called the Director Task used to assess Theory of Mind (ToM) abilities in young adults. IRT and convergent validity analyses indicated that, contrary to our hypotheses, the Director Task had an unduly narrow range of responding for measuring ToM. Furthermore, the Director Task did not correlate with other established measures of ToM. Our results suggest that the task should be used with caution when assessing a young adult population. Furthermore, since convergent validity was not established, it is uncertain what specifically the task measures. Overall, we show how IRT may serve as a useful tool in evaluating behavioural measures.

Keywords: Theory of Mind, Item Response Theory, Director Task

Introduction

Item Response Theory is an approach to assessing the psychometric properties of measures designed to measure psychological constructs such as attitudes. Modern test construction methodology suggests that simply having a range of scores on a measure is not a sufficient determinant of the psychometric properties of a test. In the current research article, we extend the use of Item Response Theory (IRT) methodology from its traditional application of evaluating personality scales and achievement to validate the effectiveness of a behavioural task, specifically, a Theory of Mind task called the Director Task.

IRT provides sample invariant information for each item at varying levels of the underlying traits or ability (Embretson & Reise, 2000; Thissen & Wainer, 2001). The simplest IRT model is the dichotomous Rasch (1960) model which is applied to tests, or other tasks, where each trial can be classified as correct or incorrect. The Rasch model allows us to calculate the difficulty of each item (1PL), its discriminatory power (2PL), as well as account for the effect of guessing (3PL). By calculating the probability of answering each question correctly based on assumed trait levels, IRT can supply researchers with information about the suitability of individual test items, as well as the test in general. IRT provides a number of advantages over classical test construction methods, such as allowing for identification of sensitivity and difficulty of individual items (Embretson, 1996; Hambleton & Swaminathan, 2013). Most crucially, IRT allows researchers to empirically assess the suitability of the test at varying levels of the trait of interest. This information allows researchers to determine the effective range of discrimination for the tool.

IRT models make four major assumptions: unidimensionality, local independence, monotonicity, and a normally distributed latent trait. Unidimensionality of the trait and local independence are generally assumed to coexist. Unidimensionality is the assumption that there is only one latent trait being measured, whereas local independence is the assumption that each response is independent and only conditional on the latent trait. Monotonicity is the assumption that as the latent trait increases, so does the probability of correctly responding to each trial. Finally, the assumption of a normally distributed latent trait is common to many parametric tests used in psychology research. To our knowledge, IRT has never been applied to data from a behavioural task. However, there are, in principle, no conceptual restrictions that would restrict the use of IRT for the assessment of a behavioural measure.

Theory of Mind (ToM) is a cognitive ability that allows individuals to mentalize about other's minds (Heider, 1958). ToM is believed to be an important component of empathy which, along with emotion empathy, allows individuals to accurately recognize and understand other's emotional states (Smith, 2006). Disruptions in ToM abilities can lead to impairments in adult functioning where individuals are less able to interpret the beliefs and intentions of others (Perner, Frith, Leslie, & Leekam, 1989). Theory of Mind deficiencies are closely associated with Autism Spectrum Disorders (Baron-Cohen, Leslie, & Frith, 1985).

Unlike emotion perception, which is largely an inborn ability and therefore, develops extremely early (Grossmann, 2010), ToM abilities continue to develop beyond childhood. For example, infants can discriminate emotional faces at 3.5 months (Montague & Walker-Andrews, 2002), or possibly earlier, and at 6.5 months are able to differentiate between emotional postures of adults (Zieber, Kangas, Hock, & Bhatt, 2014a, 2014b). In contrast, ToM skills develop much later in life (Calero, Salles, Semelman, & Sigman, 2013; Frith & Frith, 2001). ToM development is even believed to stretch into early adulthood (Dumontheil, Apperly, & Blakemore, 2010), as evidenced by the continued neurodevelopment of brain regions responsible for ToM such as the medial frontal gyrus, the anterior paracingulate, and the right temporoparietal junction (Kana, Keller, Cherkassky, Minshew, & Just, 2009) into late adolescence and early adulthood (Shaw et al., 2008).

From a practical point of view, the assessment of ToM abilities poses a particular difficulty for clinicians and researchers. Many tasks that measure the development of ToM abilities, such as the presence of false beliefs or perspective taking, have ceiling effects since these abilities are well developed by the age of 5 (Wellman, Cross, & Watson, 2001). Other measures, such as the Reading the Mind in the Eyes Task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), are confounded by the emotion perception aspect of the task. However, some perspective taking tasks, such as the Director Task (Keysar, Barr, Balin, & Brauner, 2000) have been shown to discriminate ToM abilities later into adolescence, and even early adulthood (Keysar, Lin, & Barr, 2003).

The Director Task is a perspective taking task where the participant is instructed to follow directions of a confederate who has a different view of a 4 x 4 grid. The grid contains various items that the participant must manipulate based on the director's instructions. Some of the grids are closed to the view of the director, but not the participant (see Figure 1). During the experimental trials of the task, the director gives an ambiguous instruction to the participant to move an item (e.g.: "Move the bottom block"). In this example, there would be two distractor blocks, one of which is the lower most from an egocentric perspective, but is closed off (i.e., unable to be seen) from the view of the director, and therefore is not the target. If participants select the lower-most block that is visible to them, they would not have taken the director's perspective into account, and would thus commit an error.

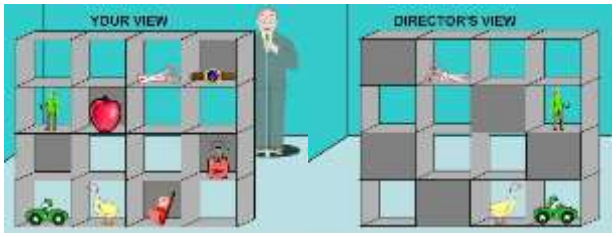

Figure 1. Instruction examples given to participants to demonstrate the director's perspective.

In previous studies, the Director Task showed that even adults have a natural tendency for the egocentric perspective (Keysar et al., 2003), and that the task reliably differentiates between youth and young adults (Dumontheil et al., 2010). These findings indicate that the Director Task may be a useful tool to differentiate between Theory of Mind abilities within the young adult/ adult population. If it is true that the task can reliably differentiate between young adults on ToM abilities, this would allow for the study of ToM perspective taking using convenience samples, making ToM research more accessible.

Present Study

The purpose of the present study was to assess the Director Task using IRT. Specifically, would the Director Task prove suitable for use with the young adult population as a tool for discriminating between individuals who are low and those who are high in Theory of Mind abilities? We hypothesize that a modified, computer based, version of the Director Task would allow for the discrimination across a sample of young adults on the basis of ToM abilities, and the results would show convergent reliability with more established measures of ToM. Although the Director task has already been shown to differentiate between age groups (Dumontheil et al., 2010), this finding does not automatically extend to within group differentiation.

With regard to convergent validity, two established ToM tasks were selected, the Reading the Mind in the Eyes Task ("Eyes Task") (Baron-Cohen et al., 2001) and the 40-item Empathy Quotient (EQ 40) (Baron-Cohen & Wheelwright, 2004). Although these tasks are sufficiently different from the Director Task, we predicted that a weak, but significant positive correlation would be observed between these tasks and the Dirtector Task.

Method

Participants

94 Carleton University undergraduate students (20 male) with a mean age of 19.8 (SD = 4.3) volunteered to participate in exchange for course credit. All participants self-identified as right-handed.

Measures

As part of a larger study participants completed the Eyes Task (Baron-Cohen et al., 2001), as well as the 40 item Empathy Quotient (Baron-Cohen & Wheelwright, 2004). The Director Task (Keysar et al., 2000) used was kindly provided by Dumontheil et al. (2010) and modified for use with PsychoPy software (Peirce, 2007). The Director Task was modified to exclude the non-director items, allowing a doubling of the number of Director trials to 95. Altogether, 16 trials were experimental trials, 16 trials were control trials, and the rest of the trials were filler trials. If our hypothesis is correct, by increasing the number of experimental trials, a grater range of scores will be observed, and with it, a finer discrimination of individuals along the latent trait associated with ToM.

Procedure

After providing informed consent, participants were tested individually in a sound-attenuated booth. Instructions, stimuli, and questionnaires were presented on a PC using PsychoPy software (Peirce, 2007). The task was presented to participants as a static image with verbal instructions played over computer speakers. For each trial the target item was overlaid by a 3 cm^2 invisible square which would record mouse button presses. All mouse presses outside of the target square were scored as incorrect; trials with no mouse button presses were discarded. Each Director Task maximum trial length was set to 5 seconds from the onset of audio instructions. Trials in the Director Task were presented to participants in a predetermined order. Next, participants completed the Eyes Task and the EQ 40 task.

Trials in these latter tasks were randomized. The study required approximately 45 minutes to complete. Participants were debriefed as to the purpose of the experiment after they completed the EQ 40 task.

Results

Responses were tallied and scored using custom Visual Basic scripts. Outliers were identified based on deviations from predicted Mahalanobis distance using the R package "careless" (Yentes & Wilhelm, 2018). One case was identified as unusual and removed leaving 93 participants (see Figure 2). Descriptive statistics are presented in Table 1. Scores from the Director Task appeared to take on a bi-modal distribution, with upper and lower scores trending towards extremes (see Figure 3).

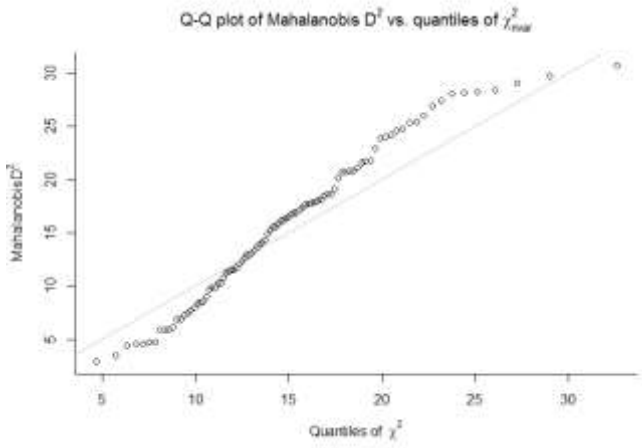

Figure 2. q-q plot of actual vs predicted Mahalanobis distance

Table 1. Overall descriptive statistics for each measure.

	M	SD
EQ 40 Score	68.21	7.80
Eyes Task	26.80%	5.39%
Director Task	53.14%	35.74%

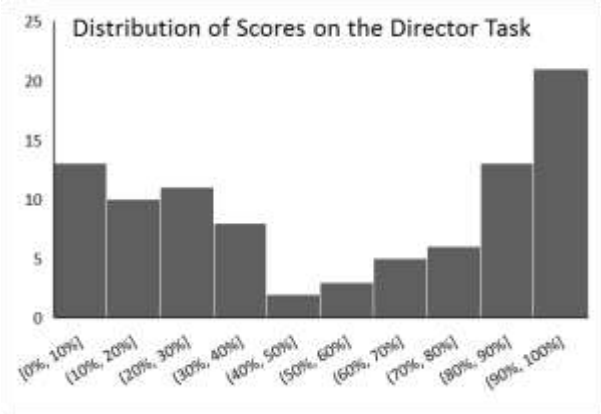

Figure 3. Histogram showing the distribution of accuracy scores on the Director Task

A paired samples t-test showed a significant decrease in accuracy when comparing the control trials with the experimental trials (t (92) = -9.05, $p < .01$) but not reaction times (t (92) = -1.14, *ns*). This suggests that performance on the task deteriorated as expected due to the increased difficulty of the experimental trials compared to the control trials.

IRT

The IRT analysis was performed using the ltm (Rizopoulos, 2006) package in the R environment (R Core Team, 2013). A constrained One-Parameter Logistic Model (1PL) and unconstrained Two-Parameter Logistic Model (2PL) dichotomous models was run to determine which created a better fit. The constrained model assumes that each item on the unidimensional scale is equally good at discriminating between individuals with varying trait levels whereas the unconstrained model does not make this assumption. Since the two models are nested, a χ^2 difference test was performed to assess model fit.

Significant model fit improvement was observed when the model was unrestricted from constrained to the unconstrained discrimination parameters (χ^2 (14) = 27.97, p = 0.014). As such, a 2PL model was selected for the analysis of the Director Task. A 3PL model was not used because the Director Task is not strictly a forced choice multiple choice test, and therefore it is improbable that participants would attempt to randomly select their answers.

Results of the individual item difficulty and discrimination, under the 2PL model, are presented in Table 2. Figure 4 contains the Item Information Curves (IIC) and Figure 5 Shows Total Test Information Function relative to Standard Error of measurement. Standard errors were estimated using the delta method.

The results from the model suggest that, congruent with our hypothesis, all the experimental trials of the Director Task have good discriminatory power. However, contrary to our hypothesis, the difficulty of the items appears quite low with only half of the items showing a significant deviation from 0.

Table 2. Difficulty and discrimination of the experimental items of the Director Task

Trial	Difficulty (*b*)	Discrimination (*a*)
4	-0.33	1.40**
14	-0.03	1.16**
20	-0.18	2.43**
26	-0.06	2.52**
30	-0.27*	2.66**
36	-0.34**	3.10**
40	-0.07	4.00**
49	-0.026	3.62**
59	-0.61**	2.22**

65	-0.40**	3.32**
70	0.17	2.25**
74	-0.67**	1.66**
78	-0.36**	3.72**
84	-0.18	2.21**
88	-0.44**	2.38**

Note: * $p < .05$; ** $p < .01$

The IIC plot visually confirms that, although the information content of many trials is very high, the range of ToM ability that they represent is poor.

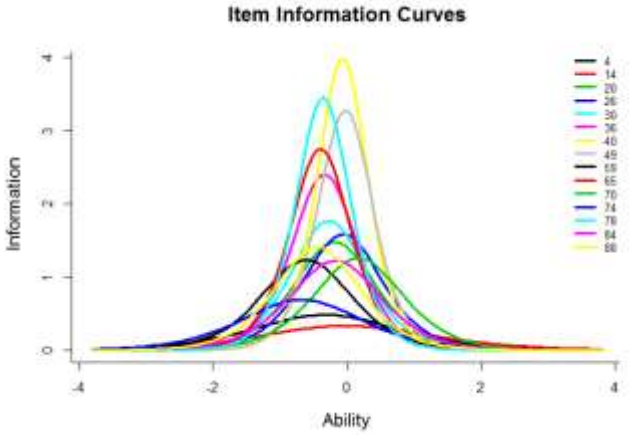

Figure 4. Item Information Curves for the experimental items of the Director Task

Finally, Figure 5 shows that the information content of the Director Task as a whole is very large, with an area under the curve of 38.66. However, 55% (20.75) of this information content falls within 0.5 standard deviations of the mean, and 82.5% (31.89) within 1 standard deviation. This once again reaffirms that the Director Task is poor at discriminating between individuals of different ToM abilities.

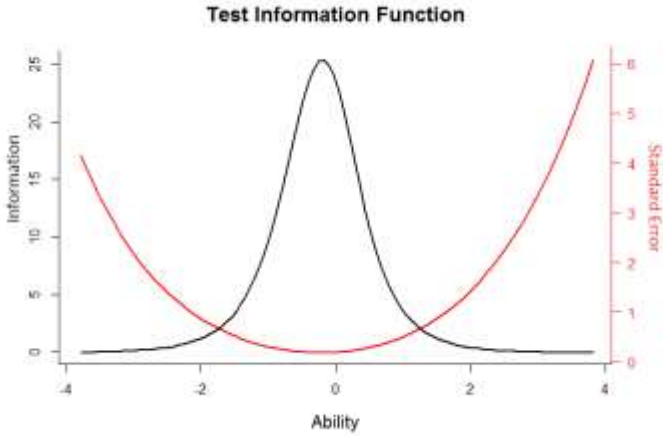

Figure 5. Total Test Information Function relative to Standard Error of Measurement.

Convergent Validity

Convergent validity for the Director Task were assessed using a self-report measure of ToM, the EQ 40, as well as a behavioural discrimination task, the Eyes Task. The results are presented in Table 3.

Table 3. Correlation matrix for the Director and other convergent validity tasks

	1	2
EQ 40	-	
The Eyes Task	-0.102	-
The Director Task	0.065	0.114

Contrary to our hypothesis, we did not find any significant correlations between the Director Task, or any of the other two popular tasks for assessing ToM abilities.

Discussion

Our findings did not support the hypothesis that the Director Task is good at discriminating between Theory of Mind abilities in a sample of young adults. Our findings are surprising in light of previous findings with the same (Dumontheil et al., 2010) or similar (Keysar et al., 2003) tasks allowing for discrimination in the young adult population.

Our sample showed significant variability in the range of scores on this task, which under normal circumstances would be an encouraging finding. However, IRT analysis showed that despite strong information content of the individual trials (discrimination), the Director Task does not measure well different levels of the ToM trait (difficulty). We interpret these findings as a strong indication that the Director Task is able to differentiate participants as either good or bad at TOM abilities, with little useful information beyond that. This interpretation is supported by both the poor difficulty gradient of the trials, as well as the tendency for participant scores to conform to a bimodal distribution.

Beyond the poor psychometric properties of the task, we failed to observe convergent validity between the Director Task and other established ToM tasks. This finding brings into question what trait or state the Director Task actually is measuring. One possible explanation for the lack of relationship between the three ToM Tasks examined in this study is that there is a sufficiently large distinction between the perspective taking ToM component, and emotion perception ToM component. However, this would not explain the lack of relationship between the Eyes Task and the EQ scores. Another possible explanation is that the Director Task is measuring some other quality, such as selective-attention to the task (Rubio-Fernández, 2017).

Regardless, we would caution researchers using the Director Task in its present form. Specifically, the task suffers from overly homogenous difficulty of trials. Nonetheless, there is potential for a modified version of this

task to be more successful. If the task is modified such that there is a greater range of experimental trial difficulty, with some being more difficult, while others being easier, the likely utility of the task will greatly improve. Finally, it is possible that by assessing other behavioural measures beyond the accuracy of answers, such as mouse-tracking or eye-tracking (Symeonidou, Dumontheil, Chow, & Breheny, 2016) we could use the extra sources of information to supplement our inferences about participants' ToM abilities.

Regarding the more general goal of extending IRT to assess the results of a behavioural task by validating the Director Task, the present results suggest that IRT can provide useful information about the relationship between participants' responses and the construction of tasks. IRT is often associated with pen and paper test construction, however, the underlying probability models are agnostic to the source of the data. With many available statistical packages, and a well developed literature, IRT is easily accessible to all researchers. We encourage the use of IRT as a readily available tool to aid the validation of measures.

Conclusion

In this study we used Item Response Theory to validate the Director Task (Keysar et al., 2000) as a tool in studying Theory of Mind abilities in young adults. Contrary to our hypotheses, we found that the task performed poorly in discriminating between levels of the latent trait. Furthermore, a convergent validity measure brought into question what latent trait is being measured using the Director Task. Overall, the present study provided a novel demonstration of how an Item Response Theory analysis can be profitably extended to assess behavioural measures.

References

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition, 21*(1), 37-46. doi: 10.1016/0010-0277(85)90022-8

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders, 34*(2), 163-175. doi: 10.1023/b:jadd.0000022607.19833.00

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241-251. doi: 10.1017/S0021963001006643

Calero, C.I., Salles, A., Semelman, M., & Sigman, M. (2013). Age and gender dependent development of theory of mind in 6-to 8-years old children. *FRONTIERS IN HUMAN NEUROSCIENCE, 7*(May), 281. doi: 10.3389/fnhum.2013.00281

Dumontheil, I., Apperly, I.A., & Blakemore, S.-J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science, 13*(2), 331-338. doi: 10.1111/j.1467-7687.2009.00888.x

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science, 10*(5), 151-155. doi: 10.1111/1467-8721.00137

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience, 28*(2), 219-236. doi: 10.3233/RNN-2010-0499

Hambleton, R.K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*: Springer Science & Business Media.

Heider, F. (1958). *The psychology of interpersonal relations*. New York Wiley.

Kana, R.K., Keller, T.A., Cherkassky, V.L., Minshew, N.J., & Just, M.A. (2009). Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Social Neuroscience, 4*(2), 135-152. doi: 10.1080/17470910802198510

Keysar, B., Barr, D.J., Balin, J.A., & Brauner, J.S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32-38. doi: 10.1111/1467-9280.00211

Keysar, B., Lin, S., & Barr, D.J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25-41.

Montague, D.P.F., & Walker-Andrews, A.S. (2002). Mothers, fathers, and infants: The role of person familiarity and parental involvement in infants perception of emotion expressions. *Child development, 73*(5), 1339-1352. doi: 10.1111/1467-8624.00475

Perner, J., Frith, U., Leslie, A.M., & Leekam, S.R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child development*, 689-700.

R Core Team. (2013). R: A language and environment for statistical computing.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Rizopoulos, D. (2006). Ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5).

Rubio-Fernández, P. (2017). The director task: A test of theory-of-mind use or selective attention?

Psychonomic Bulletin & Review, 24(4), 1121-1128. doi: 10.3758/s13423-016-1190-7

Shaw, P., Kabani, N.J., Lerch, J.P., Eckstrand, K., Lenroot, R., Gogtay, N., . . . Rapoport, J.L. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *Journal of Neuroscience, 28*(14), 3586-3594.

Smith, A. (2006). Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record, 56*(1), 3-21.

Symeonidou, I., Dumontheil, I., Chow, W.-Y., & Breheny, R. (2016). Development of online use of theory of mind during adolescence: An eye-tracking study. *Journal of Experimental Child Psychology, 149*, 81-97. doi: 10.1016/j.jecp.2015.11.007

Thissen, D., & Wainer, H. (2001). *Test scoring*: Routledge.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development, 72*(3), 655-684.

Yentes, R.D., & Wilhelm, F. (2018). Careless: Procedures for computing indices of careless responding (Version 1.1.3): R package

Zieber, N., Kangas, A., Hock, A., & Bhatt, R.S. (2014a). The development of intermodal emotion perception from bodies and voices. *Journal of Experimental Child Psychology, 126*, 68-79. doi: 10.1016/j.jecp.2014.03.005

Zieber, N., Kangas, A., Hock, A., & Bhatt, R.S. (2014b). Infants' perception of emotion from body movements. *Child development, 85*(2), 675-684. doi: 10.1111/cdev.12134