Child Abuse & Neglect 117 (2021) 105062

Contents lists available at ScienceDirect

Child Abuse & Neglect

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

# Appropriate responses to potential child abuse: The importance of information quality

Annalese Bolton [a,b,c,*], Simon Gandevia [c,d], Ben R. Newell [e]

[a] UNSW Forensic Psychology Clinic, UNSW, Sydney, Australia
[b] Matilda Centre, The University of Sydney, Sydney, Australia
[c] Neuroscience Research Australia (NeuRA), Sydney, Australia
[d] School of Medicine, UNSW, Sydney, Australia
[e] School of Psychology, UNSW, Sydney, Australia

ARTICLE INFO

*Keywords:*
Potential child abuse
Reporting
Health/allied health
Decision Making
Child welfare
Training

ABSTRACT

*Background:* When people suspect a child or young person is experiencing, or at risk of, abuse or neglect, they have to decide how to respond. However, the under and over reporting of child welfare issues indicate that people may struggle to identify an appropriate response.
*Objective:* To develop scenarios (for future training and research purposes) that closely resemble the child welfare situations health/allied health practitioners encounter, and for which there is a reasonable level of child protection professional consensus as to what the appropriate response for each situation should be.
*Participants, setting, and methods:* We developed 285 scenarios from 190 child protection reports made by health/allied health practitioners to two Australian government child welfare agencies, that covered a range of appropriate response pathways and abuse types. An appropriate response pathway for each scenario was identified by having 34 child protection professionals provide their opinion and rationales.
*Results:* Child protection professionals displayed moderate (e.g., krippendorf's alpha = 0.58, 95 % CI: 0.52 to 0.62) interrater agreement as to the appropriate response pathway for the scenarios. For 127 of the 285 scenarios (44.56 %), there was strong consensus ($K = 0.73$, 95 % CI: 0.66 to 0.78).
*Conclusion:* Professional consensus was higher than anticipated from previous research, although still low compared to generally acceptable levels of consensus. Our results suggest several promising avenues to increase professional consensus, such as improving the quality of information that people typically report to child welfare agencies.

## 1. Introduction

All over the world, people who come in contact with families have to decide how to respond if they come across a child or young person they suspect is at risk of abuse or neglect. Traditionally, this decision focused on 'whether or not' to make a report to the relevant statutory child protection agency. However, globally, many child protection systems have shifted away from a forensic-focused, dichotomous 'child protection or not' view, towards a differential response approach (Butchart, Harvey, Mian, & Fürniss,

* Corresponding author at: School of Psychology, University of New South Wales, UNSW, Sydney, NSW, 2052, Australia.
*E-mail address:* a.bolton@unsw.edu.au (A. Bolton).

https://doi.org/10.1016/j.chiabu.2021.105062
Received 13 October 2020; Received in revised form 16 February 2021; Accepted 30 March 2021
Available online 8 April 2021

2006; Merkel-Holguin, Kaplan, & Kwak, 2006).

Differential response systems (sometimes also referred to as 'public health model', 'alternative', 'family assessment', or 'multi track' response systems; e.g., Butchart et al., 2006; English, Wingard, Marshall, Orme, & Orme, 2000; Fuller, Paceley, & Schreiber, 2015; Hughes, Rycus, Saunders-Adams, Hugues, & Hugues, 2013; Scott, Lonne, & Higgins, 2016) place greater emphasis on early detection and intervention support for vulnerable families. Although there are variations in the way differential response systems are implemented across countries and jurisdictions, the same central ideas underpin these systems (Hughes et al., 2013; Kaplan & Merkel-Holguin, 2008). Such as the idea that families should be empowered to address the issues themselves whenever possible, and the intrusiveness of the response required should vary depending on the severity of risks to the child. The most intrusive statutory responses are therefore limited to situations where the risk of harm to the child or young person is considered unacceptably high (Merkel-Holguin et al., 2006). In differential response systems, available services have expanded to address the different levels of risk severity, and the response options available have widened considerably. For example, Table 1 displays the different categories of response pathways available in New South Wales (NSW), Australia.[1]

As outlined in Table 1, in NSW's system, the Child Protection Helpline functions as a reporting body and gateway for statutory involvement for situations established as meeting the *Risk of Significant Harm* (*ROSH*) threshold. The Helpline does not typically provide referrals, support, or have further involvement with families that are deemed below the *ROSH* threshold. To support the vulnerable 'at-risk' families who fall below the *ROSH* threshold, Child Wellbeing Units (set up within government health, education and law enforcement departments) and various local non-government Early Intervention Services were established. This means that when people have concerns about a child's welfare, they have to decide what type of response pathway is most appropriate.

However, the extent to which people who interact with families – such as police and law enforcement, teachers and school staff, health practitioners, social service professionals, family and friends – can identify an appropriate response within differential response systems remains unknown. In fact, we are unaware of any prior research that explores the appropriateness of people's judgements that go beyond the dichotomous 'is it, or is it not' a child protection issue, or 'should you, or should you not' report this to child protection services (e.g., Rodriguez, 2002).

### 1.1. The definition of an appropriate response

We operationalise an 'appropriate response' in terms of decision alignment. Particularly, we want to align people's decisions about what response pathway to initiate for any given situation with the expectations of the child protection system they are in. For instance, can people accurately identify situations that require a statutory child protection service? Can people accurately identify situations that Early Intervention Service staff would deem to be appropriate referrals?

Our operational definition is different to what is typically assumed about normative decision-making research within child protection. Typically, it is assumed that normative definitions require child and family outcomes (e.g., López, Fluke, Benbenishty, & Knorth, 2015). However, defining appropriate responses by child and family outcomes is problematic as outcome information is often not available. When it is available, it is difficult to determine if it reflects the quality of the initial decision due to a multitude of confounds, such as the type and effectiveness of interventions that were, or were not, provided as a result of the initial decision (López et al., 2015). Our definition, on the other hand, side steps the issue of whether a decision was ultimately 'right or wrong'. Rather, our definition allows us to explore how best to assist people to effectively use the child protection system they are in to get the most appropriate support for the vulnerable children, young people, and their families that they come in contact with.

### 1.2. Inaccurate responding: under and over reporting

Unfortunately, people's responses do not always align with the expectations placed on them by child protection frameworks. Sometimes people do not make reports to statutory child protection services when they should. For example, prior to Christoffer Kihle Gjerstad's death in 2005, Christoffer had presented to Norwegian health services multiple times for injuries caused by severe violence. Despite being aware of Christoffer's injuries, no one from his family, school, or health services notified child protection services or the police (Sethi et al., 2013). The non-reporting of abuse may not be uncommon (e.g., Flaherty et al., 2008; Sege & Flaherty, 2008). For example, Flaherty et al. (2008) had paediatric clinicians collect data prospectively for child injury visits. Clinicians recorded information about the injury, child, family, likelihood that the injury was caused by child abuse, and whether the injury was reported to child protective services. The clinicians suspected that ~10 % of 1683 injuries (recorded by 327 clinicians) were caused by abuse, however only 6% resulted in a report to child protection services. Twenty seven percent of the injuries considered 'likely' or 'very likely' caused by abuse, were not reported.

There are multiple reasons why suspected abuse is not reported (e.g., see reviews McTavish et al., 2017; Sege & Flaherty, 2008). One of the primary reasons was a belief that alternative responses (e.g., on-going support and involvement with the family, treatment, etc.) would more effectively address the issues than involving child protection services (e.g., Jones et al., 2008). This reason fits with differential response systems, and the response that is often expected of people who come in contact with families.

On the other hand, emerging evidence suggests that people continue to rely on reporting to statutory services when alternative responses are more appropriate (Cassells et al., 2014). For example, approximately half of the reports made to the NSW Child

[1] For an extended version of this Table, see Appendix A.

**Table 1**
Risk Response Pathway Categories.

| Not Child Protection | Problematic | Heightened Risk/ Need (*HRN*) | (Suspected) Risk Of Significant Harm (*ROSH*) | Immediate Response (*iROSH*) |
|---|---|---|---|---|
| **Definition of child /young person** | **Category definition** | **Category definition** | **Category definition** | **Category definition** |
| Young people under the age of 18 years are covered by Australian child protection legislation. | The parents / caregivers have behaved, or there is a risk that they might behave or fail to behave, in a way that has caused, or may cause, **minor harm to a child / young person in their care.** | The parents/caregivers have, or there is a risk that they might, behave or fail to behave in way that has, or is likely to, cause **harm to a child / young person.** | The parents/caregivers have, or there is a reasonable risk that they might, behave or fail to behave in way that has, or is likely to, cause **significant harm to a child / young person.** | The parents/caregivers have, or there is a reasonable risk that they might, behave or fail to behave in a way that has, or is likely to cause, **imminent harm to a child / young person.** |
| **Most invasive response required:** | **Most invasive response required:** | **Most invasive response required:** | **Most invasive response required:** | **Most invasive response required:** |
| No child protection response required. | Raise your concern with the family, support the family to address the issue/s, document and monitor. | Support the family to get involved with Early Intervention Family Support Services. | Make a report to the Statutory Child Protection Helpline. | Immediate report to the Child Protection Helpline, and Call Emergency Services (e.g. Police, Ambulance) if needed. |

This table was developed by AB with contributions from NSW Health Child Wellbeing Unit and Child Protection Helpline staff.
*Note.* Each column represents a response category option.

Protection Helpline do not meet the *Risk of Significant Harm* threshold for statutory intervention (Bolton, Newell, Gandevia, Peek, & Berrocal Capdevilla, 2019; Cassells et al., 2014).

The Family and Community Services Behavioural Insights Unit interviewed people who had made reports to the Child Protection Helpline, and found that many were unclear as to what makes a situation reach the *Risk of Significant Harm* threshold and that many felt that the child protection training they had received did not equip them to identify when an alternative response would be more appropriate (Family & Community Services Behavioural Insights Unit., 2016). The need for people to receive higher quality child welfare training is not unique to NSW. Similar findings are reported in other Australian states and all over the world, such as the United States, United Kingdom, Sweden, Norway, Taiwan, El Salvador, and Malta (e.g., Bjørknes, Iversen, Nordrehaug Åstrøm, & Vaksdal Brattabø, 2019; Borg & Barlow, 2018; Christian, 2008; Feng, Chen, Wilk, Yang, & Fetzer, 2009; Flaherty, Jones, Sege, & Child Abuse Recognition Experience Study Research Group, 2004; Foster, Olson-Dorff, Reiland, & Budzak-Garza, 2017; Gilbert et al., 2009; Goldman, 2007; Herendeen, Blevins, Anson, & Smith, 2014; Hurtado, Katz, Ciro, & Guttfreund, 2013; Talsma, Boström, & Östberg, 2015).

### *1.3. The lack of empirical research as to what makes child protection training effective*

Unfortunately, there is a paucity of empirical research on what makes child protection training effective. A worldwide systematic review identified only 15 child protection training studies (Carter, Bannon, Limbert, Docherty, & Barlow, 2006). Of these, only four used an objective outcome measure, only three used a control group, and all evaluated multiple and confounding interventions in a single study design. Since this review, rigorous empirical research with objective outcome measures has remained rare. What is available consists mostly of program evaluations (e.g., Alvarez et al., 2010; Cerezo & Pons-Salvador, 2004; Kenny, 2007; Leppäkoski, Rantanen, Helminen, & Paavilainen, 2019; Mathews et al., 2017; Paranal, Washington Thomas, & Derrick, 2012; Smeekens et al., 2011), rather than investigations of what does or does not improve the appropriateness of response decision-making.

### *1.4. Aim and scope of our overall research program*

The first aim of our research program is to examine *if* and *where* the response pathway judgements of people who come in contact with families (e.g., members of the general public, people whose work brings them in contact with children or young people, etc.) diverge from the response pathways identified by child protection professionals. The second aim is to explore how we can accurately and efficiently address any identified divergence using insights from cognitive and educational psychology (e.g., Newell, Lagnado, & Shanks, 2015). We plan to capitalise on our knowledge of how humans think, learn, remember and make optimal decisions to identify potential training elements, and then test how effective these elements are at improving people's ability to identify appropriate responses. By identifying effective training elements, we can hopefully improve the quality of child protection training programs across the world. However, to conduct this research, we first needed to develop ecologically valid materials.

### *1.5. Purpose and aim of the research outlined in this paper*

This paper details how we developed a set of 285 scenarios based on real situations that varied across the response pathway categories outlined in Table 1 and across suspected abuse types. For each scenario, staff from two government child welfare agencies (The Child Protection Helpline and Health Child Wellbeing Unit) provided their professional opinions and rationales as to which of the response pathways in Table 1 was most appropriate.

The type of child welfare situations experienced by people who come in contact with families and the information they have access to varies substantially depending on the person's role and level of involvement with the family. Therefore, we focused on the type of situations experienced by just one group of people: health and allied health practitioners – such as medical, dental, and general health practitioners, nurses, midwives, psychologists, and occupational therapists.

This paper is laid out in two sections. First, we detail how we developed the scenarios, and then we expound on how the scenarios were validated against the response pathways in Table 1.

## 2. Scenario development

Ethics approval was obtained through the South Eastern Sydney Local Health District's and the University of New South Wales' Human Research Ethics Committees (approval numbers: HREC#17/261 and HREC#/17/POWH/603 respectively) to develop the scenarios.

We partnered with the New South Wales Child Protection Helpline, and the New South Wales Health Child Wellbeing Unit to develop the scenarios. The Child Protection Helpline's primary focus is to identify situations where the risks of harm to the child or young person are so significant that an invasive statutory response is required, such as situations that meet the *Risk of Significant Harm* (*ROSH*), or *Immediate Risk of Significant Harm* (*iROSH*) thresholds as outlined in Table 1. The Health Child Wellbeing Unit provides support for situations below the *ROSH* threshold, such as situations that fall in the *Problematic* or *Heightened Risk/Needs* categories outlined in Table 1.

These government agencies provided 190 'typical' child protection reports made to their organisation by health/allied health practitioners. These reports varied in the severity of risk to the child or young person, and ranged from situations where agency staff had determined that no child welfare response was required (i.e., situations that fell in the *Not Child Protection* response pathway category in Table 1) through to situations of such high risk that an immediate statutory response was required (i.e., situations that fell in the *iROSH* category in Table 1). The reports also spread across a range of primary suspected abuse types as outlined in Table 2. Before providing us with these reports, the agency staff removed all identifiable information about the health/allied health reporter and the families on which the reports were based.

Using these 190 real reports, we created 285 scenarios. To do this, we first identified the important features to guide the type of information to be included in each scenario. These features are outlined in Table 3 and were identified from three sources. First, we abstracted key themes from the evidence-based structured decision-making tools used in NSW's child protection system - the Mandatory Reporter Guide (MRG) and the Screening and Response Priority tool (SCRPT). Both these tools are part of the Structured Decision Making suite of evidence-based assessment tools and decision guidelines developed by the National Council on Crime and Delinquency Children's Research Centre in California, North America (Freitag & Park, 2008) to support and guide practitioners' decision-making in relation to child welfare situations. These tools are used in various jurisdictions in North America and across the world. Second, we identified key features from our own practical experience making response recommendations (e.g., the family's past and current supports, and the family's attitudes towards and engagement with support services). Third, other key features were identified through consultation with the Director of the Health Child Wellbeing Unit (e.g., household occupants).

Next, in consultation with staff from the government agencies we partnered with, we developed a structure that was applied to each scenario (see Figs. 1 and 2). To further reduce the possibility that the reporter or subjects of the original reports could be identified, when creating each scenario, we altered specific details (e.g., ages, family structure, medical diagnoses, locations, genders, ethnic background). Vague or abstract details were replaced with concrete descriptions. When details were changed, added, or removed, precautions were taken to ensure the scenario remained realistic. For example, if a medical diagnosis was changed, relevant literature was reviewed using PubMed and/or we consulted with health/allied health practitioners to ensure the details of the new diagnosis were medically sound. Attempts were made to include both positive (i.e., strengths, resources, protective factors) and negative factors (i.e., problems, concerns, risk factors\indicators) in each scenario (as suggested by Turnell & Edwards, 1999).

The original reports rarely contained information concerning services that the family was already engaged with, their attitudes towards, or engagement with such services. To provide realistic information relating to these features we conducted searches for appropriate services, and where appropriate, spoke to service representatives (e.g., methadone clinics, non-government organisations, etc.).

To increase the likelihood that the final set of scenarios would include a reasonable spread across the response pathway categories outlined in Table 1, we used the two decision tools used within the jurisdiction (i.e., the MRG and SCRPT tools). These tools contain guidance as to what particular response pathways should be initiated when certain features are present in the situation. We used this specific information to make variations of some of the scenarios to alter the likely recommended response pathway outcome.

To do this, we first identified a primary abuse type for each scenario as defined by the decision tools (see Table 2 for the abuse type definitions). The decision tools were then applied to the scenarios by two forensic psychology Masters students plus one of the authors (AB) to identify a likely recommended response pathway. Differences of opinion were discussed to derive agreement as to the primary suspected abuse type and likely response pathway. To make variations of some scenarios, we then used the decision tools to identify which features of the scenario we could change to alter the likely recommended response pathway. Furthermore, we also included scenarios where we had purposefully removed/not included information about crucial features identified in the MRG and SCRPT tools, which we earmarked as potentially falling into an additional category labelled '*Insufficient Information*'.

The MRG and SCRPT do not have the same labels for recommended outcomes as used in Table 1. Table 4 outlines the alignment between the MRG and SCRPT tool outcomes with the response categories in Table 1.

Finally, the scenarios were proofread by two additional researchers. The entire set of scenarios cannot be made publicly available

**Table 2**
Suspected Abuse Types, and Alignment with the MRG Decision Trees.

| Suspected Abuse type | Broad definitions [a] | MRG decision trees |
|---|---|---|
| **Carer concerns** | Child/young person appears to be significantly affected by carers substance abuse, mental health, or domestic violence, or the child/young person is in voluntary care for longer than legislation allows. | Carer concerns: substance abuse, mental health, and domestic violence. Relinquishing care. |
| **Prenatal** | There is concern about the welfare of an unborn child at birth. | Unborn child. |
| **Neglect** | There is suspicion that a parent/carer is not adequately meeting the child/young person's needs (e.g. supervision, shelter, medical care, hygiene/clothing, mental health care, schooling/education, nutrition, other basic needs). | Neglect: supervision, physical shelter/environment, food, hygiene/clothing, medical care, mental health care, and education (not enrolled and habitual absence). |
| **Physical** | There is suspicion that a non-accidental injury or physical harm to a child/young person may have been caused by a parent/carer or other adult household member. | Physical abuse. |
| **Sexual** | Sexual activity or behaviour that has been imposed, or is likely to be imposed, on a child/young person by another. | Sexual abuse: of a child (age 0–15 years), of a young person (age 16–17 years), child / young person displays problematic sexual behaviour towards others. |
| **Emotional *b*** | A child/young person appears to be experiencing emotional/psychological distress, or is a danger to self or others, as a result of parent/carer behaviour such as domestic violence, carer's mental health, carer's substance abuse, or there is an underage marriage or similar union that has occurred or is being planned. | Psychological harm. |
| **Child / Young person is a danger to self / others** | A child/young person is danger to self and/or others (C/YP Danger to S/O) and it is unclear whether the parent/carer behaviours contributed now or in the past. | Child/young person is danger to self and/or others. |

*Note.* [a] These broad definitions were derived from the MRG (National Council on Crime & Delinquency, 2016) – see https://reporter.childstory.nsw.gov.au/s/mrg for more details. [b] The terms emotional abuse and psychological harm are often used interchangeably.

**Table 3**
Situational Features.

| Feature | Main question | Additional questions |
|---|---|---|
| Harm | What is the harm I am concerned about or suspicious of? | What is the source of suspicion? E.g. For suspected sexual abuse, has the child/young person been exposed to sexually inappropriate stimuli or behaviour? |
| Likelihood | How likely is it that this harm will occur in the future? | |
| Immediacy | How soon might this harm occur? | e.g. Does the alleged perpetrator still have access to the child/young person? When is the child/young person likely to see the alleged perpetrator next? |
| Impact | What impact is the issue of concern having, or is likely to have, on the child/young person? | e.g. Physical injury, safety, impacts to child/young person's social, emotional/psychological, behavioural, or academic functioning, or development of life skills? |
| Severity | How severe is the harm/impact on the child/young person? | e.g. Is it life threatening? To what extent will it impact their functioning and well-being in the short and long term? |
| Vulnerability | Is the child/young person particularly vulnerable? | e.g. Are they an infant or below 5 years of age? Do they have medical, developmental, or emotional/behavioural issues? etc. |
| Caregiver contribution | How is the caregiver contributing to, or failing to address, the harm to the child/young person? | Are they directly causing, contributing to, or exacerbating the harm by omission (neglect/failing to act) or commission (abuse)? |
| Caregiver risk indicators | What caregiver risk indicators are present? | e.g. Is there a volatile home environment? Is there domestic violence? Do the caregivers have their own significant abuse/neglect history, punitive attitudes towards discipline, substance abuse issues, severe mental health disorder? For prenatal concerns - how far through the pregnancy are they? Is there a High-Risk Birth Alert (HRBA) on the system? |
| Household occupants | Who resides in the home with the child/young person, and what is their role? | Are they protective or do they pose additional risks? Have any been perpetrators of serious prior neglect or abuse (e.g. previous children removed, caused death of a child) or violence offences? |
| Child/young person's opinion | What does the child/young person say about the situation? | Do they fear for their own safety? Are they afraid to go home? Can they identify someone in their family they feel safe with? What would they like to see happen to address the problem? |
| Supports | What formal (services) and informal (i.e. family, friends, community networks) supports does the family have? | Is the family already working with services to address the issue of concern? Does DCJ or a Non-Government Organization have an open case plan with the family? How long have these services been involved? What is the family's attitude towards and engagement with these services? |
| Attitudes and Engagement | What is the child/young person's, caregiver's, and other household occupants' attitude towards and engagement with support services? | Is the family making progress to address the issues? |

**Health Professional's Title and Workplace:** You are a nurse within a hospital.

**Family Details (all household members, names, birth dates, and contact details):** The mother gave birth 4 days ago at 38 weeks gestation. This is the mother's first child. The unnamed newborn is currently in the intensive care unit (ICU), as he has bradycardia and will require a pacemaker and on-going check-ups throughout his development. The mother stated that the father was "not involved".

**Family's Relationship to Your Service (include relevant information from a review of your files):** Your organisation's files indicate that this is the first time the mother has attended your hospital.

**Current Concerns (including additional information available to you, e.g. via Chapter 16A[a]):** You are concerned with the mother's ability to organise and attend regular health check-ups for her baby once she is discharged. The mother had no antenatal care and claimed she only found out she was pregnant 10 weeks ago, but "did not get around" to seeing anyone. Her behaviour towards her baby is loving, but she is easily distracted, disorganised, frequently loses track of time (e.g. twice, she has gone for lunch and not been back in time to meet with doctors, and she often starts feeding her infant just before scheduled appointments) and is poor at following through with simple tasks (e.g. she has repeatedly stated that she wants to express milk but does not "get around" to it).

**Services, Support, Engagement and Progress:** You are not aware of any services or supports that the mother is engaging in. The mother has not had any visitors. The mother's family live inter-state and she says that they are not able to help.

**Fig. 1.** Scenario Example One.
*Note.* [a] Chapter 16A of the NSW *Children and Young Persons (Care and Protection) Act 1998* allows information to be exchanged between prescribed bodies (without statutory child protection involvement) to facilitate the provision of services to children, young persons, and their families when there are concerns about a child or young person's welfare. This information can be shared despite other laws that prohibit or restrict the disclosure of personal information, such as the *Privacy and Personal Information Protection Act 1998*, the *Health Records and Information Privacy Act 2002* and the *Commonwealth Privacy Act 1988*. This is because the needs and interests of children and young persons, and of their families, in receiving services relating to the care and protection of children or young people takes precedence over the protection of confidentiality or of an individual's privacy.

due to the sensitive nature of the material[2] and to further guard against potential identification of vulnerable families and the practitioners who made the original reports. However, a couple of the more generic examples are given in Figs. 1 and 2.

## 3. Validating the scenarios against the response categories

To validate our scenarios against the response pathway categories outlined in Table 1, staff from both the Child Protection Helpline and Health Child Wellbeing Unit provided their opinion and rationales for their response pathway category choice for each scenario. By having staff from two government child welfare agencies rate the scenarios, we have a sound understanding of what the appropriate response for each situation should be in a child welfare system as a whole, as opposed to the type of response given by any agency in isolation.

Appropriate response pathways were determined by the majority consensus of these child protection staff. Child welfare situations are highly complex, and even among staff in the same system, there is variation and inconsistent assessment and decision-making (Barber, Shlonsky, Black, Goodman, & Trocmé, 2008; Bartelink, Van Yperen, Ten Berge, De Kwaadsteniet, & Witteman, 2014; Kang & Poertner, 2006; Levi & Crowell, 2011; Stokes & Schmidt, 2012), let alone variations between organisations (Keddell, 2014; Shlonsky & Benbenishty, 2013).

Evidence-based structured decision-making tools, such as the MRG (used at Health Child Wellbeing Unit) and SCRPT (used at the Child Protection Helpline), reduce inconsistencies (e.g., Johnson, Wagner, & Wiebush, 2000). However, the extent to which they do so is unclear, as we were unable to obtain any inter-rater reliability estimates for either the MRG or SCRPT, or about the consistency between the two tools.

## 4. Methodology

Approval was obtained through the South Eastern Sydney Local Health District's and the University of New South Wales' Human Research Ethics Committees (approval numbers: HREC17/284 and HREC/17/POWH/604 respectively) to conduct the following study

[2] Although the scenarios may be made available on request through the corresponding author.

**Health Professional's Title and Workplace:** You are a dentist at a clinic.

**Family Details (all household members, names, birth dates, and contact details):** 6-year-old Mary and her 10-year-old brother Aaron live with their birth mother (single-parent household).

**Family's Relationship to Your Service (include relevant information from a review of your files):** This is the first time the family has attended your service.

**Current Concerns (including additional information available to you, e.g. via Chapter 16A):** You are concerned about the mother's ability to manage her children's behaviour, which places them at risk of ongoing severe dental issues. Mary presented with a slim build but appeared otherwise healthy and well-nourished. During an oral examination of the children, you identified that both children needed to have multiple teeth removed due to tooth decay. Aaron told you that he and Mary eat a lot of junk food (lollies and chips) and do not brush their teeth, even though their mother "screams" at them to. You ruled out other potential causes for the tooth decay. When you raised your concerns with their mother, the mother appeared defensive and angry, and told you that she knows about appropriate dental hygiene, but that she has "given up" trying to get her children to eat healthy food and brush their teeth because they "don't listen" to her.

**Services, Support, Engagement and Progress:** The mother told you that she currently does not have any support with managing her children. When you raised the topic of her accessing parenting support services, the mother was defensive but said she would think about it.

**Fig. 2.** Scenario Example Two.

**Table 4**
Response Pathway Category Alignment with the MRG and SCRPT outcomes.

| Response category | MRG outcome | SCRPT outcome |
|---|---|---|
| *Immediate Risk Of Significant Harm(iROSH)* | Immediate report to the Child Protection Helpline | *ROSH* 24 h |
| Suspected *Risk Of Significant Harm (ROSH)* | Report to the Child Protection Helpline | *ROSH* 72 h or 1 week |
| *Heightened Risk/Needs* | Child Wellbeing Unit/Consult | Non-*ROSH* |
| *Problematic* | Document and Monitor | Non-*ROSH* |
| *Not Child Protection* | – | – |

to validate the scenarios.

### 4.1. Participants

Twenty-four staff from the Health Child Wellbeing Unit and nine staff from the Child Protection Helpline provided informed consent to participate as part of their work duties. In addition, the author's (AB) judgments as to the likely response pathway were included in the data. Therefore, a total of 34 child protection professionals' judgements were included in the data. However, due to changes in work situations and priorities, fourteen participants were unable to complete the full set of scenarios allocated to them. Although, their permission was obtained to retain their partial data.

Participating staff were primarily female (27 females, 7 males), and had a tertiary degree (17 had a tertiary degree, 14 had a higher degree, and 3 had a diploma) in either social work (n = 14), social science (n = 8), psychology (n = 6), education (n = 2), health science / Aboriginal health and community (n = 2), child protection / drug and alcohol / work health safety (n = 1) and arts (n = 1). Most had worked in their current role for 1–10 years (range 6 months - 20 years) and had worked in the child protection field for 6–20 years (range 6 months - 25 years).

### 4.2. Design

We used a randomized block design as displayed in Fig. 3 to ensure that we had a roughly equal number of opinions from each government agency and for each scenario. The scenarios were randomly and equally divided into blocks, and participants were randomly allocated to one of the blocks. To prevent decision fatigue from impacting any scenario, the presentation order of scenarios in each block was randomized for each participant.

The government agencies had each allocated 100 h of their staff's time to participate in the study. Due to differences in

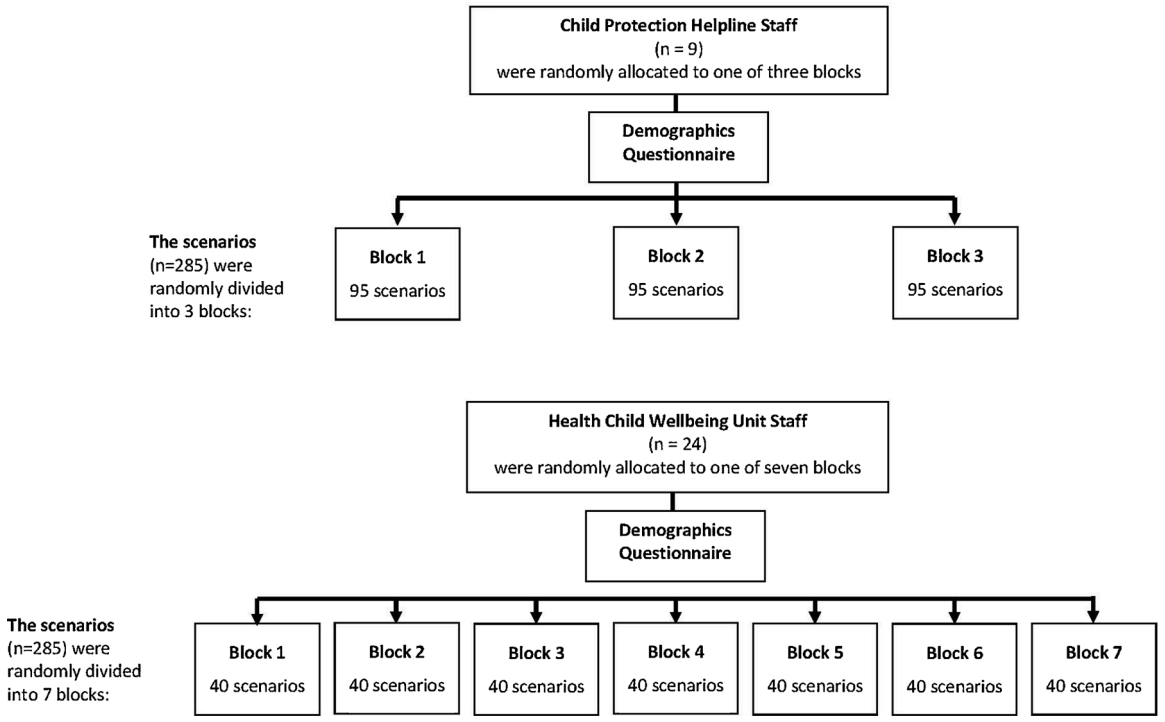


**Fig. 3.** Flow Diagram.

participation rates between the two agencies, the number of blocks and scenarios per block varied between the agencies. The scenarios for Health Child Wellbeing Unit staff were divided into seven blocks of 40 scenarios, whereas the scenarios for the Child Protection Helpline staff were divided into three blocks of 95 scenarios.

### *4.3. Procedure and materials*

Participants first read information about the purpose of the study and an explanation of how the study would proceed. They were provided with an extended version of the response pathway categories outlined in Table 1, information about how the response pathway categories in Table 1 fit with the decision tool outcomes their agency used, and information about the important features to consider. Next, they were provided with two worked examples of how to provide their judgement and opinions for a scenario (see Appendix A and B for the information participants were provided).

Participants then completed an online survey run via the Qualtrics platform. This survey had previously been pilot tested by representatives from the Child Protection Helpline, Health Child Wellbeing Unit, academics, and higher degree research students. The survey remained open to participants for four months enabling progressive completion, at times arranged with their managers.

The online survey included a demographic questionnaire at the start followed by the scenarios. The scenarios (e.g., Figs. 1 and 2) were presented one at a time. For each scenario, participants made a judgment as to which response pathway category outlined in Table 1 they would recommend. They then provided their rationale for a) "Choosing this category (e.g., what were the important features that made you select this category?)", b) "Not choosing a *higher* category" (except when they chose *iROSH*), and c) "Not choosing a *lower* category" (except when they chose the *Not Child Protection* category). They also provided their recommendation as to what the health practitioner in the scenario should do next. Finally, participants were given an opportunity to flag whether the scenario contained any unrealistic information and to identify what that information was.

If participants did not feel that there was enough information in the scenario to confidently recommend a response pathway category, they had the option of choosing an additional category called "*Insufficient Information*". If the participant indicated that the scenario contained *Insufficient Information* they were prompted to identify what crucial information they felt was missing.

On each page, before they provided their responses, participants were reminded of the purpose and value of their responses (i.e., "We value your expertise and we want to use your knowledge to help [people respond appropriately to child welfare situations]. Therefore, the quality of your responses to the following questions are very important").

### *4.4. Data analyses*

To determine an appropriate response pathway for each scenario, we plotted the number of staff who chose each response pathway category for each scenario, and then divided the scenarios into the agreements' levels as defined in Table 5. An appropriate response

**Table 5**
Criteria for the Scenario Agreement Level.

| Agreement level | Criteria for agreement level | Number of Scenarios (% of Total 285) |
|---|---|---|
| 100 % | All participants chose same response pathway category. | 26 (9%) |
| Strong | When the number of participant opinions for the scenario was n ≤5, the appropriate response pathway category had two or more votes than any other category. When n ≥6, the appropriate response pathway category had three or more votes than any other category. | 101 (35 %) |
| Majority | When the number of participant opinions for the scenario was n ≤5 the appropriate response pathway category had one more vote than any other category; When n ≥6 the appropriate response pathway category had two votes more than any other category. | 105 (37 %) |
| Boundary | There were two neighbouring response pathway categories where there was an equal, or nearly equal split between participants as to the appropriate category. | 36 (13 %) |
| No consensus | Scenarios where none of the above criteria applied. | 17 (6%) |

was only appointed for scenarios where there was *at least* a majority consensus.

Participants' level of agreement was further assessed with two measures. First, we calculated a proportion of agreement score ($P_a$). To do this, we divided the number of ratings for the appointed appropriate response category by the total number of ratings (this included all the ratings for scenarios where there was no consensus).

Second, we used a weighted Kappa ($K$) using Krippendorff's alpha (Hayes & Krippendorff, 2007; See Watson & Petrie, 2010, for why one would choose this approach over other statistical approaches), using the R-function 'kra' from the package 'rel'. The number of simulations was set at R = 1000 (as recommended by Efron, 1992) to calculate bootstrap confidence intervals (CI). Krippendorff's alpha is a measure of inter-rater agreement for ordinal data (with at least three categories) where there are varying numbers of coders (e.g., participants) per item (e.g., scenarios), provided that there are at least two coders per item (i.e., a minimum of two ratings per item). One of the assumptions of Krippendorff's alpha is that there is a reasonable distribution of items across the categories. It corrects for the likelihood that coders randomly select the same response option for an item, and it is weighted so that the extent of disagreement is considered. For example, a difference of one category (e.g., *iROSH* to *ROSH*) is represented as less disagreement than a difference of three categories (e.g., *iROSH* to *Problematic*).

$K$ scores range from -1 to 1. A negative $K$ indicates that raters disagree more than would be expected by chance, $K = 0$ indicates chance agreement, positive scores indicate agreement, and $K = 1$ indicates perfect agreement (Fleiss, Nee, & Landis, 1979). There is no definitive threshold for an acceptable level of agreement, however, Landis and Koch (1977) suggest the following: $K = 0.00$ to $0.20$ = 'slight' agreement, $K = 0.21$ to $0.40$ = 'fair' agreement, $K = 0.41$ to $0.60$ = 'moderate' agreement, $K = 0.61$ to $0.80$ = 'substantial' agreement, and $K = 0.81$ to $1.00$ = 'almost perfect' agreement.

Few empirical studies have investigated the level of consensus when it comes to identifying appropriate response pathways for suspected child abuse or neglect situations, although what is available suggests that the inter-rater agreement levels are rarely above $K \geq 0.5$ (a moderate level of agreement at best). For instance, Bartelink et al. (2014) investigated the level of agreement of child protection professionals in the Netherlands who were trained on a structured decision-making tool ($n = 40$ & $n = 33$) versus those who were not trained on the tool ($n = 40$ & $n = 31$). They found little evidence of a significant difference between the two groups, with $K$'s of 0.24 and 0.09 for those trained on the tool and $K$'s of 0.06 and 0.26 for those who were not trained. Kang and Poertner (2006) investigated the inter-rater reliability of 45 caseworkers using the Illinois structured decision support protocol for three suspected child abuse scenarios. They found a $K$ of 0.29. Baird, Wagner, Healy, and Johnson (1999) compared the inter-reliability of nine child protection caseworkers, recruited from across the United States of America, and trained to use one of three decision tools – the Washington Risk Assessment Matrix, the California Family Assessment Factor Analysis, or the Michigan Family Risk Assessment of Abuse and Neglect. They found $K$s of 0.18, 0.18, and 0.56 respectively using a sample of 80 scenarios.

The $K$s recorded in these previous studies provide inter-rater reliability estimates for specific decision tools rather than agreement across multiple organisations using different decision tools, such as in our study. Nevertheless, these previous studies serve as an approximate guide in the current context. For instance, it is reasonable to expect that the overall $K$ for our scenarios would be below $K \geq 0.5$.

### 4.5. Required sample size for Krippendorff's alpha

Determining a minimum sample size (i.e., number of ratings = the number of coders x the number of items) for $K$ statistics is important because when the sample size is too low, the proportion of agreement (i.e., $P_a$) can be high, but the $K$ is unexpectedly low (e.g., Feinstein & Cicchetti, 1990; Krippendorff, 2012). Under such circumstances, $K$ is unreliable. However, the point where this occurs depends on several parameters. For example, smaller sample sizes are required when $K$ is larger, the number of categories is greater, confidence intervals are calculated, the number of ratings per item is greater, the number of coders or items is greater, and when ordinal categories and weighted $K$s are used (e.g., Bujang & Baharum, 2017; Shoukri, Asyali, & Donner, 2004).

Currently there is no way to calculate the sample size required for Krippendorff's alpha. To determine an approximation of the required sample size, we performed sample size calculations for Cohen's kappa (i.e., agreement between two raters) with the R-function 'N.cohen.kappa' in the 'irr' package. To do this, we set the value of the kappa under the null hypothesis as $K = 0$, and the minimum acceptable level of agreement as $K = 0.2$ and $K = 0.4$ respectively. For 80 % power (at alpha = 0.05, two-sided), with two

ratings and six categories per item, the minimum sample size to detect $K \geq 0.2$ and 0.4 is 219 and 56 ratings respectively. Given that we have greater than two ratings per item and are using a weighted statistic, this may be an overly conservative estimate.

## 5. Results

For each scenario, between four to eight child protection staff provided their ratings, and most scenarios had five to seven ratings each (i.e., there were four child protection staff opinions for 33 scenarios, five for 94, six for 93, seven for 56, and eight for 9 scenarios). The total sample size was $N = 1624$ ratings. Examination of the plots (see Appendix C for examples of these plots) depicting participants' response pathway judgements for each scenario by the participants' place of employment revealed no obvious pattern of difference between staff from the Child Protection Helpline versus Child Wellbeing Unit in terms of their response category choices.

### 5.1. Participants' overall level of agreement

The Krippendorff's alpha indicated that for the entire set of scenarios, there was 'moderate' agreement between child protection staff as to the appropriate response category,

$P_a = 0.54$ and $K = 0.58$ (95 % CI: 0.52 to 0.62).

Examination of the scenarios by the agreement levels (as outlined in Table 5) revealed that the child protection staff displayed *at least* a *Strong* agreement for 127 (44 %) of the scenarios. These included scenarios from across all the response pathway categories and abuse types (Table 6). For this subset of scenarios, the Krippendorff's alpha indicated that there was 'substantial' agreement, with $P_a = 0.78$ and $K = 0.73$ (95 % CI: 0.66 to 0.78).

There was a clear *Majority* consensus for 105 (37 %) scenarios, and for 36 (13 %) of the scenarios, child protection staff were divided between two closely related response pathway categories (Tables 5 and 6). For 17 scenarios, there was no clear consensus at all.

### 5.2. Participants' level of agreement by response pathway category

For scenarios where an appropriate response category had been identified (i.e., where there was at least a *Majority* agreement), there was a greater proportion of agreement (i.e., $P_a$) for scenarios where the appropriate response pathway required a statutory response (i.e., the *iROSH* and *ROSH* category response pathways) or where no child protection response was required (i.e., the *Not Child Protection* response pathway; Table 7). The proportion of agreement was lowest for situations that required primarily low-level support and ongoing monitoring (i.e., the *Problematic* response pathway) or assistance to engage with early intervention services (i.e., the *Heightened Risk Needs* response pathway).

**Table 6**
Number of Scenarios by Agreement for Appropriate Response Pathway and Abuse Type.

| | | Agreement Level | | | |
|---|---|---|---|---|---|
| | 100 % | Strong | Majority | Boundary | No consensus |
| **Appropriate response category** | | | | | |
| *Not Child Protection* | 5 | 18 | 9 | – | – |
| *Problematic* | 0 | 11 | 18 | – | – |
| *Heightened Risk/Needs* (HRN) | 0 | 12 | 25 | – | – |
| *Risk Of Significant Harm* (ROSH) | 7 | 34 | 21 | – | – |
| *Immediate Risk Of Significant Harm* (iROSH) | 14 | 16 | 21 | – | – |
| *Insufficient Information* | 0 | 10 | 11 | – | – |
| *Total (% of all scenarios)* | *26 (9%)* | *101 (*35 %*)* | *105 (37 %)* | | |
| **Boundary Scenarios** *a* | | | | | |
| *Not Child Protection / Problematic* | – | – | – | 17 | – |
| *Problematic/HRN* | – | – | – | 21 | – |
| *HRN/ROSH* | – | – | – | 23 | – |
| *ROSH/iROSH* | – | – | – | 28 | – |
| **Abuse type** | | | | | |
| Carer concerns | 2 | 26 | 33 | 11 | 7 |
| Prenatal | 1 | 12 | 11 | 3 | 4 |
| Neglect | 8 | 20 | 17 | 6 | 1 |
| Physical | 3 | 17 | 12 | 4 | 0 |
| Sexual | 6 | 9 | 10 | 1 | 1 |
| Emotional | 3 | 4 | 13 | 4 | 4 |
| Child /Young Person Danger to Self/Others | 3 | 7 | 9 | 4 | 3 |
| Unclassified | 0 | 6 | 0 | 0 | 0 |

*Note.* [a] Some scenarios are counted both in the *Majority* agreement level and as a *Boundary Scenario*.

**Table 7**
Participants' Response Pathway Judgments by the Identified Appropriate Response Pathway.

| | Participants' Response Pathway Judgements | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Appropriate Response Pathway | *Insufficient Information* | *Not Child Protection* | *Problematic* | *Heightened Risk /Needs* | *Risk of Significant Harm* | *Immediate Risk of Significant Harm* | Total | $P_a$ |
| *Insufficient Information* | 74 | 10 | 16 | 13 | 6 | 1 | 120 | ***0.62*** |
| *Not Child Protection* | 10 | 125 | 22 | 8 | 8 | 2 | 175 | ***0.71*** |
| *Problematic* | 5 | 25 | 98 | 33 | 10 | 1 | 172 | ***0.57*** |
| *Heightened Risk /Needs* | 13 | 10 | 29 | 125 | 32 | 7 | 216 | ***0.58*** |
| *Risk Of Significant Harm* | 19 | 2 | 16 | 34 | 238 | 40 | 349 | ***0.68*** |
| *Immediate Risk Of Significant Harm* | 9 | 4 | 4 | 13 | 48 | 216 | 294 | ***0.73*** |
| *No Consensus* | 34 | 40 | 61 | 60 | 74 | 29 | 298 | |

**Table 8**
Inter-Rater Reliability (Krippendorff's alpha) by Abuse Type.

| Abuse type | $P_a$ | *K* | (95 % CI) | No. of scenarios | No. of ratings |
|---|---|---|---|---|---|
| Child /Young Person danger to Self /Others | 0.51 | 0.67 | (0.45−0.79) | 26 | 159 |
| Neglect | 0.58 | 0.61 | (0.47−0.72) | 52 | 301 |
| Sexual | 0.66 | 0.57 | (0.31−0.74) | 27 | 154 |
| Physical | 0.60 | 0.57 | (0.37−0.72) | 36 | 203 |
| Prenatal | 0.50 | 0.45 | (0.29−0.59) | 31 | 177 |
| Emotional | 0.45 | 0.45 | (0.24−0.61) | 28 | 156 |
| Carer concerns | 0.48 | 0.42 | (0.30−0.52) | 79 | 442 |
| Unclassified [a] | 0.75 | 0.27[b] | (-0.02−0.52) | 6 | 32 |

*Note.* [a] These scenarios consisted of three *Not Child Protection*, two *Insufficient Information*, and one *iROSH* (classified as *iROSH* for the sole reason that out-of-hours consent from the Child Protection Helpline was required for an urgent medical procedure for a child under the care of the minister). [b] Given that the assumption of a reasonable spread across the categories is violated and there is a large discrepancy with the $P_a$, the low number of scenarios and relatively low number of ratings, it is likely that the *K* estimate for the Unclassified abuse type is unreliable.

**Table 9**
Appropriate Response Pathways for Scenarios Initially Identified as Insufficient Informatio.

| Consensus level | Final Appropriate Response Pathway for Scenarios Initially Identified as Insufficient Information | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Not Child Protection* | *Problematic* | *Heightened Risk /Needs* | *Risk Of Significant Harm* | *Immediate Risk Of Significant Harm* | *Insufficient Information*[a] | – |
| 100 % | 0 | 0 | 0 | 0 | 0 | 0 | – |
| Strong | 1 | 0 | 1 | 0 | 0 | 10 | – |
| Majority | 1 | 0 | 3 | 0 | 1 | 10 | – |
| No consensus | – | – | – | – | – | – | 8 |
| **Total** | **2** | **0** | **4** | **0** | **1** | **20** | **8** |

*Note.* [a] An additional scenario was also identified as containing *Insufficient Information* by consensus.

### *5.3. Participants' level of agreement by abuse type*

The level of agreement between child protection staff also varied across abuse types. Our participants displayed greater consensus for situations where a child or young person was a danger to themselves or others and when there was neglect, but least for situations that involved carer concerns, prenatal risk indicators, and emotional abuse (Table 8).

### *5.4. The link between insufficient information and lack of consensus*

There was a link between the scenarios previously identified by the research team as containing *Insufficient Information* and scenarios where there was a lack of consensus as to the appropriate response pathway category.

For example, thirty-five scenarios were originally identified by the researcher as containing *Insufficient Information*. The majority (60 %) of these scenarios were also identified by the consensus of child protection staff as containing *Insufficient Information* (Table 9). Although, for 20 % of scenarios originally identified by researchers as containing *Insufficient Information*, there was still a consensus as to an appropriate response. However, for 23 % of these scenarios, child protection staff displayed no consensus as to an appropriate response.

For 17 scenarios (6% of the total scenario set) there was no consensus regarding an appropriate response pathway. Of these 17 scenarios, eight (i.e., 47 %) had been previously identified by the researcher (AB) as containing *Insufficient Information*, and three (i.e.,

18 %) had been identified by a child protection staff member as containing confusing or unrealistic details.

## 6. Discussion

Our aim was to develop a set of scenarios that closely resembled the type of child welfare situations that health and allied health practitioners encounter, and for which there was child protection professional consensus as to an appropriate response pathway for each scenario. For 127 of the 285 scenarios we developed, there was a *Strong* consensus between child protection professionals as to the appropriate response pathway. Within this sub-set of scenarios, there was a reasonable spread of scenarios across each available response pathway and abuse type. These scenarios will allow us to explore *where* the response pathway judgements of people who suspect a child or young person is at risk of abuse or neglect diverge from the response pathways identified by child protection professionals. We can then investigate what training elements most efficiently increase alignment.

The level of agreement between the child protection professionals for the whole set of 285 scenarios was higher than anticipated. Previous research on child welfare risk assessments indicate reasonably low agreement levels, with *k*'s ranging from 0.06 to 0.56 (Baird et al., 1999; Bartelink et al., 2014; Kang & Poertner, 2006). Of these, the highest agreement rate ($k = 0.56$) was for a specific tool within a single organization (Baird et al., 1999). Evidence indicates that even within an organization, practitioners display inconsistent assessment and decision-making (Barber et al., 2008; Bartelink et al., 2014; Kang & Poertner, 2006; Levi & Crowell, 2011; Stokes & Schmidt, 2012), which is likely further exacerbated by procedural differences between organisations (Keddell, 2014; Shlonsky & Benbenishty, 2013). Yet, despite the fact that we used child protection professionals from two organizations using different decision tools, in our study, $K = 0.58$, a rate slightly higher than all previous findings.

Potentially the interrater agreement was higher than anticipated in our study because we used a more sensitive statistic (i.e., a weighted kappa for ordinal data is more sensitive to the extent of disagreement than the dichotomous agree/disagree approach used in previous research). However, given the proportion of agreement (i.e., $P_a$ value) was also high, the kappa measurement method alone cannot fully account for the higher than expected level of agreement.

Alternatively, decision making is generally more effective when the data that is relied upon is of higher quality (e.g., Price & Shanks, 2008; Shiloach et al., 2010).

Our higher than expected agreement was more likely the result of the increased quality of the information in the scenarios compared to what people typically report to child welfare agencies. For instance, when converting the original reports into scenarios, we ensured that the scenarios contained concrete descriptions and realistic details, avoided jargon, provided information about the family's strengths and risks according to the features outlined in Table 3, and communicated the information clearly and in a structured manner. Moreover, we found a link between scenarios where there was *Insufficient Information* and where there was a lack of consensus. This highlights the importance and benefits of assisting people to identify, consider, and clearly communicate relevant information. For example, structured checklists for people may assist them to gather and consider relevant information. If we can improve the quality of information provided to child welfare agencies, then hopefully we can improve the quality of subsequent decision making, the consistency of child protection professionals' responses, and the appropriateness of the responses that vulnerable families receive.

For some scenarios that we had earmarked as containing "*Insufficient Information*" the child protection professionals were nevertheless in consensus in regards to the response they felt was most appropriate. Potentially this is because the absent features were not crucial, the included features were enough by themselves to trigger a response, or perhaps the absence or presence of some features is more difficult to determine (features which may or may not have significantly altered the type of response required).

Although our *K* of 0.58 in terms of the level of agreement among child protection professionals may be higher than previous research in this area, $K = 0.58$ is still far short of the generally acceptable level of agreement, which ideally should be above $K = 0.70$. This indicates that there is likely a greater degree of under or over responding to situations than there needs to be. Any time a family becomes involved in the child welfare system, there is a significant amount of stress, shame, and even trauma associated with that involvement (e.g., see Masson & Dickens, 2015). If a child protection professional deems that a child or young person is at significant risk of harm and in need of statutory intervention, when in actuality they are not, then that family suffers such stressors unnecessarily. Likewise, if a child protection professional deems that a family is not in need of service to address the risks when in fact support services are indeed warranted, then the child or young person is at risk of experiencing ongoing harm. Therefore, it is imperative that researchers further explore how to improve the consistency of decision-making by child protection professionals.

Our results indicate areas where professional consensus was stronger versus weaker, thereby elucidating areas to target. For instance, we found greater consensus between our participants for situations that did not require a child protection response (i.e., the *Not Child Protection* response pathway), or that required a statutory response (i.e., *ROSH*, and *iROSH* response pathways), compared to situations that primarily required low-level support and ongoing monitoring (i.e., the *Problematic* response pathway) and situations where early intervention services were warranted (i.e. the *Heightened Risk Needs* response pathway). This is not surprising given that the current decision tools used by our participants (e.g., the MRG and SCRPT) focus on distinguishing situations that need a more invasive statutory response (i.e., *ROSH* and *iROSH*) from those that do not (all situations below the *ROSH* threshold), and therefore result in a more defined structure for making such distinctions. A potential area for future research would be to identify the features of situations that indicate that the *Problematic* or *Heightened Risk Needs* response pathways are most appropriate, and how this information can be incorporated into the appropriate decision tools. This is particularly important for child protection systems that want to place a greater emphasis on early intervention.

In addition, consistent with previous research (e.g., Trickett, Mennen, Kim, & Sang, 2009; Vial, Assink, Stams, & van der Put, 2019), the level of agreement between our child protection professionals also varied across the abuse types. This likely occurs because some

abuse types, such as carer concerns, prenatal risk indicators, and emotional abuse, are harder to define in concrete terms or make a link between caregiver actions and child outcomes (e.g., Bromfield & Higgins, 2004; Trickett et al., 2009). This suggests that a greater level of training for these abuse types may be warranted.

The scenarios we have developed can also be used to explore other potential avenues to increase the consistency of child protection professionals' decision-making processes. For example, there was variation in the consensus level between our child protection professional participants across our set of 285 scenarios, ranging from 100 % agreement for some scenarios, to there being no identifiable consensus regarding an appropriate response for other scenarios. Identifying whether there are predictive markers of situations that result in high versus low consensus may help us facilitate circumstances that lead to overall greater consensus. For example, are there certain features, or combination of features, that when included result in greater consensus? Machine learning techniques could be applied to the current data to identify if there are any distinctive patterns between the scenarios of high versus low consensus (e.g., Rich & Gureckis, 2019).

## 7. Conclusion

This study provides a large set of ecologically valid, high-quality scenarios that can be used to examine how best to improve reporting in child welfare situations. Encouragingly, the overall level of consensus between child protection professionals for our scenarios was higher than previous research in this area, however it was still far short of generally acceptable levels of consensus.

Our study shows that we can potentially improve the consistency of responses by child welfare professionals if we a) increase the quality and type of information that people with concerns about a child or young person's welfare communicate to child welfare professionals, b) identify the features of situations that indicate that the *Problematic* or *Heightened Risk Needs* response pathways are most appropriate and incorporate this information into the decision tools, c) provide increased training for abuse types such as carer concerns, prenatal risk indicators, and emotional abuse, and d) increase our understanding of the specific features or pattern of features within situations that result in higher or lower consensus. Improving the quality of information and decision-making at the earliest point that a vulnerable family has contact with the child welfare system has the potential to improve the quality of all subsequent decision-making for that family. Such research is crucial if we are to address the problem of inconsistent decision-making within child welfare systems. Until such research is undertaken there will be families who continue to experience unnecessary and potentially traumatic intrusion into their family, and children and young people who will be left without adequate protection.

## Declaration of Competing Interest

We have no conflict of interest to disclose.

## Acknowledgements

This research has been undertaken in collaboration with the New South Wales (NSW) Department of Communities and Justice, Community Services, and the NSW Health Child Wellbeing Unit. However, the information and views contained in this study do not necessarily, or at all, reflect the views or information held by the NSW Government, the Minister for Communities and Justice or Health or the Departments.

We thank UNSW's Stats Central consultant Ben Maslen for assisting with data analysis, and Naomi Cameron and Anita Trinh for assistance in developing the scenarios, and Bronte Montgomery-Farrer for proof-reading the scenarios.

This research did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors. Annalese Bolton was funded via the Australian Government Research Training Program Scholarship.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.chiabu.2021.105062.

## References

Alvarez, K. M., Donohue, B., Carpenter, A., Romero, V., Allen, D. N., & Cross, C. (2010). Development and preliminary evaluation of a training method to assist professionals in reporting suspected child maltreatment. *Child Maltreatment, 15*(3), 211–218. https://doi.org/10.1177/1077559510365535

Baird, C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk assessment in child protective services: Consensus and actuarial model reliability. *Child Welfare, 78*(6), 723–748.

Barber, J. G., Shlonsky, A., Black, T., Goodman, D., & Trocmé, N. (2008). Reliability and predictive validity of a consensus-based risk assessment tool. *Journal of Public Child Welfare, 2*(2), 173–195. https://doi.org/10.1080/15548730802312701

Bartelink, C., Van Yperen, T., Ten Berge, I., De Kwaadsteniet, L., & Witteman, C. (2014). Agreement on child maltreatment decisions: A nonrandomized study on the effects of structured decision-making. *Child & Youth Care Forum, 43*(5), 639–654. https://doi.org/10.1007/s10566-014-9259-9

Bjørknes, R., Iversen, A. C., Nordrehaug Åstrøm, A., & Vaksdal Brattabø, I. (2019). Why are they reluctant to report? A study of the barriers to reporting to child welfare services among public dental healthcare personnel. *Health & Social Care in the Community, 27*(4), 871–879. https://doi.org/10.1111/hsc.12703

Bolton, A., Newell, B. R., Gandevia, S., Peek, J., & Berrocal Capdevilla, E. (2019). Applying behavioural insights to child protection: Venturing beyond the low-hanging fruit. *Behavioural Public Policy*, 1–25. https://doi.org/10.1017/bpp.2019.12

Borg, K., & Barlow, J. (2018). The behaviours and perceptions of paediatricians in Malta relating to child protection work: National and international implications of a mixed-methods study. *Child Abuse Review, 27*(6), 446–467. https://doi.org/10.1002/car.2532

Bromfield, L. M., & Higgins, D. J. (2004). The limitations of using statutory child protection data for research into child maltreatment. *Australian Social Work, 57*(1), 19–30. https://doi.org/10.1111/j.0312-407X.2003.t01-1-00110.x

Bujang, M. A., & Baharum, N. (2017). Guidelines of the minimum sample size requirements for kappa agreement test. *Epidemiology, Biostatistics, and Public Health, 14*(2), 1–10. https://doi.org/10.2427/12267

Butchart, A., Harvey, A., Mian, M., & Fürniss, T. (2006). *Preventing child maltreatment: A guide to taking action and generating evidence*. Geneva: World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/43499/9241594365_eng.pdf;jsessionid=D9FB68469B848C0989ED0EEA55D582AD?sequence=1.

Carter, Y. H., Bannon, M. J., Limbert, C., Docherty, A., & Barlow, J. (2006). Improving child protection: A systematic review of training and procedural interventions. *Archives of Disease in Childhood, 91*(9), 740–743. https://doi.org/10.1136/adc.2005.092007

Cassells, R., Cortis, N., Duncan, A., Eastman, C., Gao, G., Giuntoli, G., … Valentine, K. (2014). *Keep them safe outcomes evaluation final report*. Sydney: NSW Department of Premier and Cabinet. https://www.cese.nsw.gov.au/images/stories/PDF/Eval_Rep/Schools/Keep_Them_Safe_Outcomes_Eval_Final_Rpt_2014.pdf.

Cerezo, M. A., & Pons-Salvador, G. (2004). Improving child maltreatment detection systems: A large-scale case study involving health, social services, and school professionals. *Child Abuse & Neglect, 28*(11), 1153–1169. https://doi.org/10.1016/j.chiabu.2004.06.007

Christian, C. W. (2008). Professional education in child abuse and neglect. *Pediatrics, 122*(Supplement 1), S13–S17. https://doi.org/10.1542/peds.2008-0715f

Efron, B. (1992). Six questions raised by the bootstrap. In R. LePage, & L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 99–126). Palo Alto: John Wiley & Sons, Inc.

English, D. J., Wingard, T., Marshall, D., Orme, M., & Orme, A. (2000). Alternative responses to child protective services: Emerging issues and concerns. *Child Abuse & Neglect, 24*(3), 375–388. https://doi.org/10.1016/s0145-2134(99)00151-9

Family and Community Services Behavioural Insights Unit. (2016). *Mandatory reporters: Behavioural insights project summary report.* NSW: Family and Community Services.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Feng, J. Y., Chen, S. J., Wilk, N. C., Yang, W. P., & Fetzer, S. (2009). Kindergarten teachers' experience of reporting child abuse in Taiwan: Dancing on the edge. *Children and Youth Services Review, 31*(3), 405–409. https://doi.org/10.1016/j.childyouth.2008.09.007

Flaherty, E. G., Jones, R., Sege, R., & Child Abuse Recognition Experience Study Research Group. (2004). Telling their stories: Primary care practitioners' experience evaluating and reporting injuries caused by child abuse. *Child Abuse & Neglect, 28*(9), 939–945. https://doi.org/10.1016/j.chiabu.2004.03.013

Flaherty, E. G., Sege, R. D., Griffith, J., Price, L. L., Wasserman, R., Slora, E., … Binns, H. J. (2008). From suspicion of physical child abuse to reporting: Primary care clinician decision-making. *Pediatrics, 122*(3), 611–619. https://doi.org/10.1542/peds.2007-2311

Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin, 86*(5), 974–977. https://doi.org/10.1037/0033-2909.86.5.974

Foster, R. H., Olson-Dorff, D., Reiland, H. M., & Budzak-Garza, A. (2017). Commitment, confidence, and concerns: Assessing health care professionals' child maltreatment reporting attitudes. *Child Abuse & Neglect, 67*, 54–63. https://doi.org/10.1016/j.chiabu.2017.01.024

Freitag, R., & Park, K. (2008). *The structured decision making model: An evidenced-based approach to human services*. Retrieved from: Madison, WI: Children's Research Center https://www.ojp.gov/library/abstracts/structured-decision-making-model-evidence-based-approach-human-services.

Fuller, T., Paceley, M. S., & Schreiber, J. C. (2015). Differential response family assessments: Listening to what parents say about service helpfulness. *Child Abuse & Neglect, 39*, 7–17. https://doi.org/10.1016/j.chiabu.2014.05.010

Gilbert, R., Kemp, A., Thoburn, J., Sidebotham, P., Radford, L., Glaser, D., … MacMillan, H. L. (2009). Recognising and responding to child maltreatment. *Lancet, 373*(9658), 167–180. https://doi.org/10.1016/S0140-6736(08)61707-9

Goldman, J. D. (2007). Primary school student-teachers' knowledge and understanding of child sexual abuse and its mandatory reporting. *International Journal of Educational Research, 46*(6), 368–381. https://doi.org/10.1016/j.ijer.2007.09.002

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Herendeen, P. A., Blevins, R., Anson, E., & Smith, J. (2014). Barriers to and consequences of mandated reporting of child abuse by nurse practitioners. *Journal of Pediatric Health Care, 28*(1), e1–e7. https://doi.org/10.1016/j.pedhc.2013.06.004

Hughes, R. C., Rycus, J. S., Saunders-Adams, S. M., Hugues, L. A., & Hugues, K. N. (2013). Issues in differential response. *Research on Social Work Practice, 23*(5), 493–520. https://doi.org/10.1177/1049731512466312

Hurtado, A., Katz, C., Ciro, D., & Guttfreund, D. (2013). Teachers' knowledge, attitudes and experience in sexual abuse prevention education in El Salvador. *Global Public Health, 8*(9), 1075–1086. https://doi.org/10.1080/17441692.2013.839729

Johnson, K., Wagner, D., & Wiebush, R. (2000). *South Australia department of Family and Community Services risk assessment revalidation study*. Madison, WI: Children's Research Centre, National Council on Crime and Delinquency. https://www.nccdglobal.org/sites/default/files/publication_pdf/so_aus_2000_risk_reval.pdf.

Jones, R., Flaherty, E. G., Binns, H. J., Price, L. L., Slora, E., Abney, D., … Sege, R. D. (2008). Clinicians' description of factors influencing their reporting of suspected child abuse: Report of the child abuse reporting experience study research group. *Pediatrics, 122*(2), 259–266. https://doi.org/10.1542/peds.2007-2312

Kang, H. A., & Poertner, J. (2006). Inter-rater reliability of the Illinois structured decision support protocol. *Child Abuse & Neglect, 30*(6), 679–689. https://doi.org/10.1016/j.chiabu.2005.12.004

Kaplan, C., & Merkel-Holguin, L. (2008). Another look at the national study on differential response in child welfare. *Protecting Children, 23*, 5–21. Retrieved from: https://www.ojp.gov/library/abstracts/another-look-national-study-differential-response-child-welfare.

Keddell, E. (2014). Current debates on variability in child welfare decision-making: A selected literature review. *Social Sciences, 3*(4), 916–940. https://doi.org/10.3390/socsci3040916

Kenny, M. C. (2007). Web-based training in child maltreatment for future mandated reporters. *Child Abuse & Neglect, 31*(6), 671–678. https://doi.org/10.1016/j.chiabu.2006.12.008

Krippendorff, K. (2012). A dissenting view on so-called paradoxes of reliability coefficients. In C. T. Salmon (Ed.), *Communication yearbook 36* (pp. 481–499). New York: Routledge.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374. https://doi.org/10.2307/2529786

Leppäkoski, T., Rantanen, H., Helminen, M., & Paavilainen, E. (2019). How training impacts the identification and discussion of the risks of child maltreatment: A Finnish follow-up study. *Global Journal of Health Science & Nursing, 2*, 115–125.

Levi, B. H., & Crowell, K. (2011). Child abuse experts disagree about the threshold for mandated reporting. *Clinical Pediatrics, 50*(4), 321–329. https://doi.org/10.1177/0009922810389170

López, M., Fluke, J. D., Benbenishty, R., & Knorth, E. J. (2015). Commentary on decision-making and judgments in child maltreatment prevention and response: An overview. *Child Abuse & Neglect, 49*, 1–11. https://doi.org/10.1016/j.chiabu.2015.08.013

Masson, J., & Dickens, J. (2015). Protecting unborn and newborn babies. *Child Abuse Review, 24*(2), 107–119. https://doi.org/10.1002/car.2344

Mathews, B., Yang, C., Lehman, E. B., Mincemoyer, C., Verdiglione, N., & Levi, B. H. (2017). Educating early childhood care and education providers to improve knowledge and attitudes about reporting child maltreatment: A randomized controlled trial. *PloS One, 12*(5), Article e0177777. https://doi.org/10.1371/journal.pone.0177777

McTavish, J. R., Kimber, M., Devries, K., Colombini, M., MacGregor, J. C., Wathen, C. N., … MacMillan, H. L. (2017). Mandated reporters' experiences with reporting child maltreatment: A meta-synthesis of qualitative studies. *BMJ Open, 7*(10), Article e013942. https://doi.org/10.1136/bmjopen-2016-013942

Merkel-Holguin, L., Kaplan, C., & Kwak, A. (2006). *National study on differential response in child welfare*. Retrieved from. Washington, DC: American Humane Association and Child Welfare League of America http://www.americanhumane.org/assets/docs/protecting-children/PC-DR-national-study2006.pdf.

National Council on Crime and Delinquency. (2016). *Mandatory reporter guide* (seventh edition). New South Wales Government. https://intranet.nnswlhd.health.nsw.gov.au/child-protection/wp-login.php?redirect_to=%2Fchild-protection%2Fwp-content%2Fuploads%2Fsites%2F14%2F2017%2F11%2FNSW-Mandatory-Reporter-Guide-MRG-Edition-7-_-Final-Published-00000002.pdf.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2015). *Straight choices: The psychology of decision making*. Psychology Press.

Paranal, R., Washington Thomas, K., & Derrick, C. (2012). Utilizing online training for child sexual abuse prevention: Benefits and limitations. *Journal of Child Sexual Abuse, 21*(5), 507–520. https://doi.org/10.1080/10538712.2012.697106

Price, R., & Shanks, G. (2008). Data quality and decision making. In F. Burstein, & C. W. Holsapple (Eds.), *Handbook on decision support systems 1* (pp. 65–82). Springer.

Rich, A. S., & Gureckis, T. M. (2019). Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence, 1*(4), 174–180. https://doi.org/10.1038/s42256-019-0038-z

Rodriguez, C. M. (2002). Professionals' attitudes and accuracy on child abuse reporting decisions in New Zealand. *Journal of Interpersonal Violence, 17*(3), 320–342. https://doi.org/10.1177/0886260502017003006

Scott, D., Lonne, B., & Higgins, D. (2016). Public health models for preventing child maltreatment: Applications from the field of injury prevention. *Trauma, Violence & Abuse, 17*(4), 408–419. https://doi.org/10.1177/1524838016658877

Sege, R. D., & Flaherty, E. G. (2008). Forty years later: Inconsistencies in reporting of child abuse. *Archives of Disease in Childhood, 93*(10), 822–824. https://doi.org/10.1136/adc.2006.100545

Sethi, D., Bellis, M., Hughes, K., Gilbert, R., Mitis, F., & Galea, G. (2013). *European report on preventing child maltreatment*. Retrieved from. Geneva: World Health Organization https://apps.who.int/iris/bitstream/handle/10665/326375/9789289000284-eng.pdf.

Shiloach, M., Frencher, S. K., Jr, Steeger, J. E., Rowell, K. S., Bartzokis, K., Tomeh, M. G., … Hall, B. L. (2010). Toward robust information: Data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *Journal of the American College of Surgeons, 210*(1), 6–16. https://doi.org/10.1016/j.jamcollsurg.2009.09.031

Shlonsky, A., & Benbenishty, R. (2013). *From evidence to outcomes in child welfare: An international reader*. Oxford University Press.

Shoukri, M. M., Asyali, M., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research, 13*(4), 251–271. https://doi.org/10.1191/0962280204sm365ra

Smeekens, A., Broekhuijsen-van Henten, D., Sittig, J. S., Russel, I., Ten Cate, O. T. J., Turner, N., … van de Putte, E. M. (2011). Successful e-learning programme on the detection of child abuse in emergency departments: A randomised controlled trial. *Archives of Disease in Childhood, 96*(4), 330–334. https://doi.org/10.1136/adc.2010.190801

Stokes, J., & Schmidt, G. (2012). Child protection decision making: A factorial analysis using case vignettes. *Social Work, 57*(1), 83–90. https://doi.org/10.1093/sw/swr007

Talsma, M., Boström, K. B., & Östberg, A. L. (2015). Facing suspected child abuse – what keeps Swedish general practitioners from reporting to child protective services? *Scandinavian Journal of Primary Health Care, 33*(1), 21–26. https://doi.org/10.3109/02813432.2015.1001941

Trickett, P. K., Mennen, F. E., Kim, K., & Sang, J. (2009). Emotional abuse in a sample of multiply maltreated, urban young adolescents: Issues of definition and identification. *Child Abuse & Neglect, 33*(1), 27–35. https://doi.org/10.1016/j.chiabu.2008.12.003

Turnell, A., & Edwards, S. (1999). Signs of Safety R child protection approach and framework: Comprehensive briefing paper. *Resolutions consultancy*. https://solihulllscp.co.uk/media/upload/fck/file/Signs%20of%20safety/Signs-of-Safety-Briefing-Paper-v3-1.pdf.

Vial, A., Assink, M., Stams, G. J. J., & van der Put, C. (2019). Safety and risk assessment in child welfare: A reliability study using multiple measures. *Journal of Child and Family Studies, 28*(12), 3533–3544. https://doi.org/10.1007/s10826-019-01536-z

Watson, P., & Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology, 73*(9), 1167–1179. https://doi.org/10.1016/j.theriogenology.2010.01.003