Received: 18 October 2023 | Revised: 5 January 2024 | Accepted: 12 January 2024

DOI: 10.1111/cfs.13146

ORIGINAL ARTICLE

CHILD & FAMILY SOCIAL WORK WILEY

Improving the accuracy of social work judgements: A proof-of-concept study of a training programme

Dr David Wilkins

CASCADE, School of Social Sciences, Cardiff University, Cardiff, Wales, UK

Correspondence
David Wilkins, CASCADE, School of Social Sciences, Cardiff University, Cardiff, Wales, UK.
Email: wilkinsd3@cardiff.ac.uk

Funding information
None to report.

Abstract

Child and family social workers routinely make professional judgements involving significant legal and moral questions (e.g. whether a child has been abused) and more 'everyday' issues (e.g. will the child be re-referred again if we close the case now?) Yet the world is capricious, and we rarely know with certainty what is going to happen in future or the likely impact of our different choices. Given the consequences of their judgements and decisions, it is imperative that social workers are provided with the best possible support. This paper reports a proof-of-concept study of a set of interventions to improve the judgemental accuracy of social workers: (i) a survey to identify respondents with above-average existing abilities, (ii) training sessions on cognitive debiasing and (iii) structured group working and (iv) three methods for aggregating individual judgements. Findings indicate that it is possible to measure the accuracy of social work judgements in relation to case-study materials and retrospective questions, while the feedback about the training was largely positive. Any future studies should aim to recruit a more diverse set of respondents, test judgemental accuracy in relation to prospective judgements and explore what types of questions would be most helpful for real-world decision-making.

KEYWORDS
accuracy, decision-making, interventions, judgement, social work, training

1 | BACKGROUND

The importance of judgement and decision-making in statutory child and family social work cannot be overstated. Social workers make 'decisions that are life changing, such as whether a child needs to be removed from a family' and 'countless decisions everyday … that are less high profile' (Taylor & Whittaker, 2018, p. 105). As important, social workers make many judgements where they are not (solely) responsible for the decision - about the likelihood of significant harm to children, the possible impact of different interventions, and the potential consequences of different choices (Taylor, 2017). Such judgements may be evaluated against a broad set of criteria (Hood et al., 2022), including accuracy (the extent to which they are corroborated by other knowledge and empirical events), adherence with good practice (the extent to which they accord with the law, professional ethics and values), consistency (the extent to which they are made similarly in similar cases), discrimination (the extent to which they are made differently in different cases), equity (the extent to which people from different demographic and socio-economic groups are treated fairly) and outcomes (the extent to which they enhance child and family wellbeing). Irrespective of the criteria, 'judgement and [decision making] in this field are highly skilled activities' (ibid, p. 7) with 'professionals working in the most challenging

of circumstances, balancing conflicting needs and views, juggling resources and making the finest of professional judgements' (Scottish Executive, 2006, p. 13). Given their obvious importance, it may be surprising that social workers rarely receive feedback on the quality of their judgements and decisions (Kirkman & Melrose, 2014, p. 18) and that little or no specific training exists (Featherston et al., 2019) or has been evaluated (Hood et al., 2022, p. 28).

Although this may be true for social work, in fields such as business, economics and politics considerable efforts have been made to find and develop ways of supporting human judgement and decision-making. In particular, the American political scientist Philip Tetlock and colleagues have shown that it is possible with a combination of targeted selection, training, group work and aggregation to significantly improve judgemental accuracy (Atanasov et al., 2020; Chang et al., 2016; Friedman et al., 2018; Karger et al., 2022; Mellers et al., 2017, 2019; Moore et al., 2017; Tetlock et al., 2014; Tetlock & Gardner, 2016). Despite this, it is important to acknowledge from the outset the vast differences that exist between business, economic and political judgements, and those made by social workers (Taylor & White, 2006). Any attempt to learn from Tetlock's research and apply the lessons for social work must be done cautiously and with a view to adaptation. Thus, this article describes a proof-of-concept study in which a set of interventions based on Tetlock's studies were tested to explore whether and how they might be feasible in social work.

1.1 | Accuracy in social work judgements

That social workers need to have 'good judgement' is axiomatic. The nature of good judgement is not, and there are various ways of defining and measuring it. As noted, accuracy is one such criteria. In a recent literature review, Hood et al. (2022) found seven studies with a focus on accuracy, which they defined as 'the extent to which decisions are corroborated by other knowledge' (p. 5), including (but not limited to) subsequent actions and events. Three of these studies were based in the UK. Dickens et al. (2007) considered differences in care rates between English local authorities (LAs) to investigate the connection between relative need and the frequency of children entering care. The writers concluded that 'it is hard to say whether there are children who need to become looked after, but who do not' (p. 607). Forrester (2008) tracked 400 referrals to social services, finding that most were quickly closed, and examined how many were re-referred due to suspected maltreatment. If the child was re-referred quickly, this might suggest the initial judgement was inaccurate. However, Forrester concluded that 'the level of accuracy for the identification of risk of serious abuse appears comparatively high' (p. 296). Finally in the UK, Farmer and Lutman (2014) examined 138 cases of children returning home from foster care. They discovered that 59% of them experienced recurrent maltreatment over a 5-year period, and 65% returned to foster care, which they interpreted as a sign of inaccurate judgement (p. 265).

Of the international studies, Cross and Casanueva (2009) examined a sample of 4000 American child maltreatment substantiation decisions. They found that caseworker judgements of risk, harm and evidential quality were associated with the final decision. They calculated a predictive accuracy of between 73.5% and 79.4% for unconfirmed maltreatment and between 77.5% and 87.2% for substantiated maltreatment. However, DePanfilis and Girvin (2005) found that only 42% of 129 similar decisions from New Jersey were accurate. The final two international studies explored the use of structured judgement and decision-making tools. In their study, Cyr et al. (2022) found that an attachment-based tool increased the accuracy of judgements about subsequent child abuse. Yet according to Gillingham and Humphreys (2010), practitioners may be tempted to 'strategically' adjust the information they enter into such tools to ensure the outcome reflects their own professional judgements.

In some of my own research, accuracy has been measured by asking respondents to judge the likelihood of future actions, events, and outcomes in relation to case vignettes, based on anonymized real-life referrals to social services in England (Table 1). Based on these vignettes, social workers on average do slightly better than you would expect by chance and about the same in relation to their own casework (Wilkins & Meindl, 2022).

TABLE 1 An example of an anonymised referral used previously in Wilkins and Meindl (2022) to evaluate the accuracy of social work judgements.

Referral received from the police regarding Poppy: 'Mother called the police at 2 am due to domestic violence incident between her and her partner whereby mother has sustained an injury to her forehead due to Mr E having thrown a CD at her. Mother is reported to have been heavily intoxicated, mother reported that the domestic violence has been going on for several years but has not been reported before. Mother refused to make a statement and was taken by police to extended family for her own safety. It is understood at the time that Mr E did not know where she was going. However, a further call to the police was made to report that Mr E had subsequently collected the mother from the address, this call was made by extended family members, and it was reported that Mr E was under the influence of alcohol. Mr E was seen by police with the children in the car, he was stopped and found to be 2× over the legal limit for alcohol and had enough cannabis on his person to suggest he is a habitual user. Mother has been given advice by the police but has not followed it. Mother is refusing to submit a formal complaint about her partner. The children did appear well kempt however the home environment was chaotic'
In response to this referral, using a scale of 0–100 (where 0 = *impossible* and 100 = *certain*), how likely are the following outcomes: • No further action • An assessment • Emergency removal into care • Something else
Within the next 6 weeks, how likely is the child to be made subject to: • No plan • A child in need plan • A child protection plan • A looked after child plan
Within the next 12 weeks, how likely is the child's father to attend an appointment with a drug and alcohol service?
Within the next 6 months, how likely is there to be another referral about this child?
Within the next 6 months, how likely is it that this child will come into care?

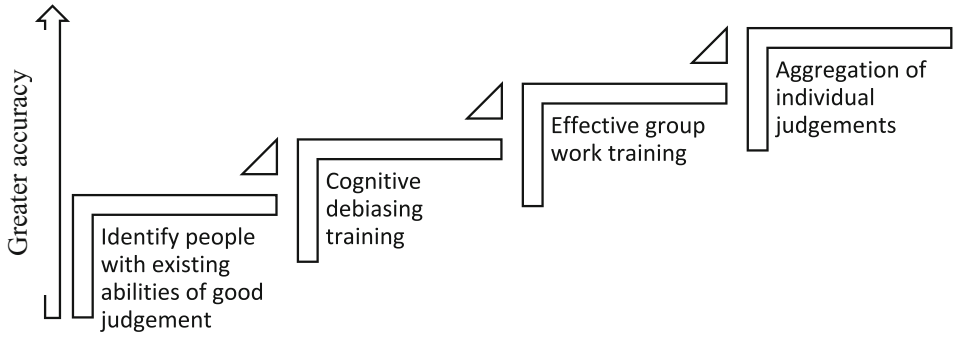

FIGURE 1 Tetlock et al.'s four-stage model for improving judgement accuracy.

It is important to note that all these studies, including my own, employ somewhat different conceptions and ways of measuring accuracy. This makes it hard to directly compare them. Notably, UK researchers tend to gather information on decisions *and* deduce the underlying judgements (e.g. Forrester, 2008), while the two American studies gathered data on judgements and decisions separately. In my work, I have so far gathered data about judgements only, and not decisions.

1.2 | How can you improve judgemental accuracy?

In their studies, Tetlock et al. have shown how it is possible to improve human judgemental accuracy under the right conditions. They do so via a four-stage set of interventions—(i) selecting people with existing attributes related to more accurate judgement, notably open-mindedness, critical thinking and pattern recognition, (ii) training in cognitive debiasing, (iii) providing a structure for effective group work and (iv) applying algorithms to aggregate individual judgements (Figure 1).

Following these methods, participants have typically outperformed the accuracy of judgements made by various comparison groups, including prediction markets (Atanasov et al., 2017) and intelligence analysts by up to 70% (Tetlock & Gardner, 2016). While the kind of questions addressed in these studies are a world-away from those that matter in social work, one can hypothesize that the same interventions could be adapted for use in social work. After all, qualities such as open-mindedness and critical thinking are already valued by the profession (Mathias, 2015), as is group work to improve critical thinking (Lietz, 2008, 2009). With greater judgemental accuracy comes not only the potential for more effective decision-making—intended actions having their intended effect (Munro, 2019b)—but a greater possibility of satisfying the moral imperative to act with fairness and proportionality, for example, balancing the child's need for protection with their parent's right to a private family life (Masson, 2006).

1.3 | Summary

Social workers routinely make judgements about children and families. The quality of which can be evaluated against various criteria, including accuracy. This is perhaps most obvious in relation to risk assessment, which inevitably involves 'predictions [about whether] current behaviour will continue … about the impact of known future events … and about an escalation of the current maltreatment' (Munro, 2019b, p. 147). However, it also applies to judgements *in general*, whenever they involve '[generating] expectations about the world and the results of our actions' (Sayer, 2010, p. 69). With this definition in mind, it is quite hard to think of a social work decision that does not (or should not) involve at least some consideration of what might or might not happen as a result. Studies of social work judgements find a range of accuracy levels, and although making consistently accurate judgements is difficult—albeit this depends on the question[1]—it is possible to make improvements. This paper reports a proof-of-concept study of Tetlock's et al.'s interventions, to explore whether they might be feasible for social work. The research questions were as follows:

1. Can Tetlock et al.'s methods be used with social workers and in relation to social work-related questions?
2. What adaptations might be needed to make them suitable for social work?

2 | METHODS

A proof-of-concept study aims to explore questions of feasibility (Schmidt, 2006). They can be helpful for determining whether a new concept or approach has potential for use in a new way, in a new context or at a larger-scale, and to identify prospective challenges and limitations. The results are often used to guide further investigation and develop more comprehensive studies in future. The interventions tested here were devised and developed entirely by Tetlock and colleagues (see www.goodjudgement.com), and no credit is being claimed for them. For this study, the interventions were undertaken as follows:

1. Identify respondents with above-average existing abilities via an online survey.
2. Sub-sample invited to attend training sessions on cognitive debiasing.
3. Training sessions on structured group working.
4. Individual judgements aggregated using three different methods (figure 2).

[1]On a scale of 0–100 (where 0 = *impossible* and 100 = *certain*), will the sun rise tomorrow?

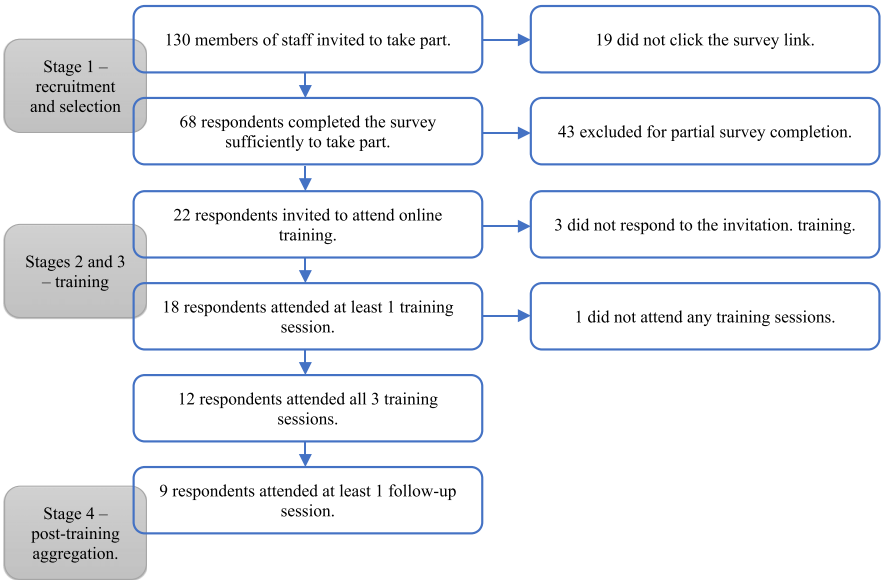

FIGURE 2 An overview of the four stages of the study.

TABLE 2 An overview of the three measures used in part one of the survey, how they are scored and how they were combined to give an overall percentage rating per respondent.

Measure description	Scoring
The Critical Reflection Test (Frederick, 2005) is designed to measure a respondent's ability to over-ride intuitive incorrect responses and engage in further reflection to find the correct response. It typically includes three items, for example: A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost? The CRT predicts the ability to make more unbiased judgements and decisions in a variety of contexts (Primi et al., 2016)	1 point per correct answer (min = 0, max = 3)
Raven's Matrices (Raven, 2008) are designed to measure a respondent's eductive ability (the ability to make sense of complexity). In our study, we used five items in which respondents were asked to identify the next sequential shape, or the missing piece from an image. While there is some debate (ibid) over the meaning of the results, it is widely believed they provide a test of general cognitive abilities (de Winter et al., 2023)	1 point per correct answer (min = 0, max = 5)
The Actively Open-minded Thinking test (Stanovich & Toplak, 2023) is designed to measure a respondent's willingness to consider alternative options, postpone closure, sensitivity to evidence contradictory to existing beliefs, and reflective thinking. In our study, we used 13 items (e.g. **people should take into consideration evidence that goes against their beliefs) and a five-point Likert-scale from strongly agree to strongly disagree. Open-minded thinking is associated with performance on most rational thinking tasks (ibid)**	Overall score obtained via a mean average (min = 1, max = 5)
For the purpose of selecting high-performing survey respondents, we combined the individual scores for each measure per respondent	Individual scores for the CRT, Raven's matrices and AOT combined and converted into a percentage. Respondents with >66% were invited to take part in stage 2

More generally, the study fits within an overall methodology of improving the quality of social work decision-making (Hilder & Whittaker, 2023).

2.1 | Stage 1: Recruiting and screening respondents

In stage 1, 130 staff in one local authority in England were invited to take part in a survey, including managers, social workers, family support workers, and administrators. The survey, hosted via Qualtrics (March 2020, Qualtrics, Provo, UT), was organized into three parts. In part one, respondents gave data about their personal and professional demographics, including sex, age, length of social care experience, role and team. Part two consisted of three standardized measures: (i) the Critical Reflection Test (Toplak et al., 2011), to measure critical reflection; (ii) a shortened version of Raven's matrices (McLeod & McCrimmon, 2021) to measure fluid intelligence or pattern recognition; and (iii) the Actively Open-minded Thinking (AOT) test (Janssen et al., 2020), to measure open-mindedness (see table 2). These measures were used to replicate the approach devised by Tetlock et al. in their studies (Mellers et al., 2015). They use them because the

characteristics measured—critical reflection, pattern recognition and open-mindedness—are associated with the ability to make more accurate judgements. For Tetlock et al., they recruit high performers in order to win forecasting tournaments (Tetlock et al., 2014). While it was not the aim of this study to win a forecasting tournament, it is reasonable nonetheless to replicate Tetlock et al.'s methods as faithfully as possible before considering what adaptations might be needed for social work.

In the third part of the survey, respondents were randomly assigned to answer questions about one of two case vignettes, based on anonymized real-life referrals to social services (referred to as Clarke and Poppy; see table 1 for an example). For each question, respondents used a numerical scale from 0 (*impossible*) to 100 (*certain*) to judge the likelihood of the specified outcome. A Brier score (Brier, 1950) was calculated for each answer, and an overall Brier score (via a mean average) for each respondent (Table 3). Brier scores provide a measure of accuracy, ranging from 0 to 2, whereby 0 indicates perfect accuracy and 2 indicates perfect inaccuracy. This was done to see whether answering questions about case vignettes might be used to provide an estimate of pre-training judgemental accuracy.

TABLE 3 Calculating a Brier score.

The formula for Brier scores is as follows, where x = the forecast for the outcome that does occur and y = the forecast for the outcome that does not occur.

$(1 - x)^2 + (0 - y)^2 = z$

Thus, if you forecast a 75% chance that a child will be re-referred within the next 6 months, and they are, your Brier score would be as follows:

$(1 - 0.75)^2 + (0 - 0.25)^2 = 0.125$

If you made the same forecast and the child was not re-referred, your Brier score would be as follows:

$(1 - 0.25)^2 + (0 - 0.75)^2 = 1.125$

Of those who completed at least the first two parts of the survey ($n = 68$), a sub-sample ($n = 22$) were invited to attend stage 2 training sessions, based on their combined score for the CRT, Raven's matrices and AOT measure. All respondents with a mean average of >66% were invited, a cut-off pre-determined on the basis that it represented a reasonable level of 'high-performance'.

2.2 | Stage 2: Training in cognitive debiasing

In stage 2, selected respondents were invited to attend two online training sessions, each lasting for 90 min. Facilitated by the author, who has previously been trained by the Good Judgement Project directly, these sessions focused on cognitive debiasing. Respondents were provided with information about common cognitive biases and a series of mitigation strategies (Table 4). The rationale is that Tetlock et al have found such training leads to a ~10% benefit in judgemental accuracy when compared to a no-training condition (Tetlock et al., 2014).

2.3 | Stage 3: Training in structured group-work

In Stage 3, the same respondents were invited to attend another online training session. This introduced respondents to the Delphi method for structured group discussions (Linstone & Turoff, 1975). Typically, this involves two rounds of discussion, in which respondents formulate their own judgements, before discussing as a group, updating their individual judgements and discussing again as a group in order to reach a conclusion (table 5). The rationale for is that Tetlock et al. have found a ~10% boost in judgemental accuracy (measured using Brier scores) via the Delphi-method, relative to individual judgements (Tetlock et al., 2014). All respondents who attended at least one of the training sessions were asked to complete a brief feedback survey.

TABLE 4 A brief overview of the cognitive bias training intervention used in this pilot study.

Cognitive bias	Mitigation strategies
Anchoring bias	Identify a suitable base-rate and amend your judgement from that starting point
Base-rate insensitivity	When forming a judgement, start with the base-rate (the outside view), before considering specific details (the inside view). Do not start the other way around (inside view first, then outside view)
Over-confidence	Participants completed a confidence calibration test, which showed they were all over-confident about their general knowledge of social work. From this, participants were asked to consider the mismatch between their subjective sense of confidence, and their real-world knowledge, and to take this into account when forming a judgement
Scope insensitivity	Participants were asked to consider practice questions resolving over different timeframes, and to notice that as the length of time increases, so does the likelihood of the outcome (and that the likelihood can never be lower for a longer timeframe compared to a shorter one)
Conjunction fallacy	Participants were given examples of how some outcomes seem more likely when they are embedded within a plausible narrative and prompted to reflect on how easy it is to be swayed by a good story
Emotional bias	Participants were asked to reflect on the lack of association between something being desirable and its likelihood (and vice versa)

Note: This is not a direct replication of the more in-depth and sophisticated training used by Tetlock et al., but our simple adaptation of it.

TABLE 5 A hypothetical example of a structured group discussion, based on the Delphi method.

An example of a group discussion, structured using the principles of the Delphi method, in relation to the following hypothetical question:

- *How likely is there to be another referral about this child in the next 6 months?*

Step 1: Each member of the group makes an individual estimation, without conferring. These are then shared with the rest of the group. For example:

- Person A: 15%
- Person B: 2%
- Person C: 25%
- Person D: 58%

(mean average: 25%)

Step 2: Each member of the group is invited to justify their own estimation and ask questions of one another about the other estimations

- Person A: I think because the current referral is not that serious, it's not that likely
- Person B: I think the parents look like they are going to engage with the assessment, so I think that will lower everyone's anxiety
- Person C: I think if you look at the history, this child has been referred before, and this is now the second time in 2 years, so on that basis I think it would be 50% in the next 12 months, and so 25% for the next 6 months.
- Person D: I think that the referrer is just very anxious, looking at their language, and so whatever happens, and given the young age of the child, I think they will make another referral again pretty soon

Step 3: After the discussion, each member of the group is asked to update their individual estimation

- Person A: 18%
- Person B: 6%
- Person C: 20%
- Person D: 30%

Step 4: The updated individual estimations are shared and aggregated (which can be achieved in various ways, the simplest being via a mean average)

- Final group judgement: 18.5%
- Assuming a negative outcome for the original question, this would result in a Brier score of 0.07, and for a positive outcome a Brier score of 1.33

2.4 | Stage 4: Post-training measurement of accuracy

Finally, in stage 4, respondents were invited to attend three follow-up sessions, facilitated by the author, during which they answered questions in relation to three case vignettes, one of which (Michael) was new to everyone and two of which (Clarke and Poppy) had been seen previously by half the respondents in the stage 1 survey. In keeping with the proof-of-concept methodology, the aim was not to ascertain the *efficacy* of the interventions, but the *feasibility* of comparing respondents' pre- and post-training accuracy using case vignettes (Table 6) and of aggregating their individual judgements using three different methods. The first involved a simple mean average (taking the sum of individual judgements and dividing by the number of respondents). The second involved an *extremizing algorithm*. This worked by increasing or decreasing the aggregated judgement of the group, depending on the direction of change between the first and second Delphi rounds (steps 1 and 3 in Table 5, above). If the mean judgement of the group went up, the algorithm increased the final judgement by 25%, to a maximum of 100%. If the mean judgement of the group went down, the algorithm decreased the final judgement by 25%, to a minimum of 0%. For the example given in Table 5, this would result in a final judgement of 13.9% (and for a negative outcome, a Brier score of 0.04). The logic is that especially when groups of people do not know each other very well, their judgements tend to congregate around 50% (Baron et al., 2014). An extremizing algorithm helps correct for this bias (Dana et al., 2019). The third method involved a *best-performer algorithm*. This was used to identify the top performers in the group and aggregate only their judgements (using a mean average). The logic is that it makes common sense to place greater weight on judgements made by respondents with the best track records (Hanea et al., 2018). For the example given in Table 5, if Persons A and B were the best performers, this would result in a final judgement of 12% (and for a negative outcome, a Brier score of 0.03). To implement this algorithm, respondents with the highest survey scores (>90%) for the CRT, Raven's matrices and AOT measures were selected.

TABLE 6 An overview of the case vignettes used in the study.

Pseudonym for the child	Brief details of the referral and example questions
1. Poppy	Female pre-school child referred by police in relation to concerns about parental alcohol misuse and domestic abuse • In the next 12 weeks, how likely is there to be another referral made about this child? • Within the next 6 months, how likely is it that the child will come into care?
2. Clark	Unborn child referred by midwife team in relation to concerns about mother's presentation and lack of engagement • In response to this referral, how likely is there to be a social work assessment? • In response to this referral, how likely is there to be no further action?
3. Michael	Male 1-year-old child, removed from the care of his mother due to concerns about substance misuse; now living with his father • Within the next 4 weeks, how likely is it that the mother will attend at least one session of contact? • In three months, how likely is it that the child will still be living with his father?

2.5 | Ethics

The study was approved by the School of Social Science ethics committee (Cardiff University) and conducted in accordance with the

British Association of Social Workers' Code of Ethics (Butler, 2002). Respondents were required to provide written consent (before the start of the stage 1 survey). All data were anonymized at the point of collection, and no identifiable information about any child or family was sought, shared or collected. Participation in the study was voluntary, and respondents were reminded at the start of each stage and training session that they could opt out without giving a reason. Any data provided were liable for inclusion in the analysis.

3 | FINDINGS

This section presents the findings from the proof-of-concept study outlined. At the outset, 130 members of staff were invited to take part, of whom 111 clicked on the survey link and 68 completed the survey sufficiently for inclusion, that is, they completed all three standardized measures (Table 7).

The aggregated results for the three standardized measures (CRT, Raven's matrices and AOT) are shown in Table 8. All respondents who provided their work email address (indicating consent to be contacted) *and* scored >66% for the combined percentage variable ($n = 22$) were invited to the stage 2 training. Of these, 18 attended at least one session, while 11 attended all three and provided feedback (Tables 9 and 10). These data indicate that Tetlock et al.'s methods for identifying individuals with above-average skills of good judgement (critical reflection, open-minded thinking and pattern recognition) can work with social workers. As can Tetlock et al.'s approach to training for cognitive debiasing and structured group work.

Following the training, the respondents who attended the follow-up sessions were able to read case vignette information and answer questions to enable the calculation of Brier scores (Tables 11–13). It was also possible to aggregate their individual judgements using a simple mean, an extremizing algorithm and a best-performer algorithm. Of these, lower Brier scores were generally achieved via the best-performer algorithm. While not the purpose of the study, these data may provide very tentative evidence that by selecting, training and combining the individual judgements made by social workers, greater levels of accuracy could be achievable.

TABLE 7 Overview of respondents' personal and professional characteristics.

Variable	Options	N	% of total
Sex	Male	7	10.3
	Female	61	89.7
	Prefer not to answer	0	0.0
Gender	Man	7	10.3
	Woman	61	89.7
	Nonbinary	0	0.0
	Other	0	0.0
	Prefer not to answer	0	0.0
Age group	18–24	5	7.4
	25–34	23	33.8
	35–44	24	25.3
	45–54	12	17.6
	55–64	4	5.9
	65+	0	0.0
Ethnicity	White (English, Welsh, Scottish, Northern Irish or British)	59	86.8
	White (any other background)	3	4.4
	Other (various options combined)	6	8.8
Social work qualified	Yes	44	64.7
	No	21	30.9
	No response	3	4.4
Social care experience	1 year or less	6	8.8
	1–3 years	13	19.1
	4–6 years	20	29.4
	7–9 years	5	7.4
	10+years	24	35.3
Team	Referral and assessment (R + A)	9	13.2
	Child in need/ child protection (CIN/CP)	18	26.5
	Edge of care	9	13.2
	Disabled children's team (DCT)	8	11.8
	Looked after children (LAC)/leaving care (LC)/fostering and adoption (F + A)	11	16.2
	Other	13	19.1

4 | LIMITATIONS

The limitations of the study include the small size of the sample and in relation to the stage 1 survey its self-selecting nature. The study was also located within only a single local authority. There is also a high risk of researcher bias, as the author designed the study, facilitated the training and follow-up sessions and collected the data. Finally, less-than-ideal comparators were used for the post-training measures of judgemental accuracy, with two of the case vignettes having been seen previously by at least some of the respondents. While these limitations are broadly acceptable from the perspective of a proof-of-concept study, they mean the findings must be interpreted with caution. Altogether, this limits the ability to make claims about whether and to what extent the accuracy of social work judgements can or should be improved using these interventions. Importantly, no such claims are being made.

5 | DISCUSSION

That the interventions explored in this study can result in significant improvements in judgemental accuracy cannot be in any reasonable doubt (Mellers et al., 2017; Mellers et al., 2019). The question is

TABLE 8 Overview of survey respondents' scores for the three standardized measures.

	N	Min.	Max.	Mean	Std. deviation
Raven's matrices overall	68	0.00	5.00	3.3088	1.32999
AOT overall	68	2.89	4.78	4.0637	0.38169
CRT overall	68	0.00	3.00	0.8824	1.19113
Valid *N* (listwise)	68				

TABLE 9 Details of the 18 respondents who attended at least one training session.

	Personal/professional characteristics					Standardized measures				Training session		
ID	Sex	Age	SW	Team	Exp.	CRT	AOT	Raven's matrices	Combined (%)	1	2	3
1	M	25–34	Y	Other	4–6	3	4.86	5	98.9	Y	Y	Y
2	F	25–34	Y	CIN/CP	1–3	3	4.44	5	95.7	Y	N	N
3	F	35–44	N	CIN/CP	4–6	3	3.78	5	90.6	Y	Y	Y
4	F	18–24	N	DCT	1–3	3	4.56	3	81.2	Y	Y	Y
5	M	25–34	Y	Other	1–3	3	4.44	4	88.0	Y	Y	Y
6	F	25–34	Y	Other	4–6	3	4.00	4	84.6	N	Y	Y
7	F	35–44	N	Other	4–6	3	4.33	3	79.5	Y	Y	Y
8	F	25–34	Y	CIN/CP	1–3	2	4.22	4	78.6	Y	Y	Y
9	M	35–44	Y	CIN/CP	10+	1	4.67	5	82.1	Y	Y	Y
10	F	45–54	Y	CIN/CP	10+	1	4.44	4	72.6	N	Y	Y
11	F	25–34	Y	CIN/CP	1–3	1	4.78	3	67.5	Y	Y	N
12	F	45–54	Y	Other	10+	1	4.44	4	72.6	Y	N	N
13	F	25–34	Y	R + A	4–6	2	4.33	3	71.8	Y	Y	Y
14	F	35–44	N	Other	4–6	3	4.00	4	84.6	Y	Y	Y
15	F	55–64	N	CIN/CP	4–6	2	4.11	4	77.8	Y	Y	N
16	F	45–54	Y	LAC	10+	1	4.67	3	66.7	Y	Y	N
17	F	25–34	Y	R + A	1–3	2	4.44	3	72.6	Y	Y	Y
18	F	18–24	N	CIN/CP	<1	1	4.86	3	68.2	Y	Y	Y
Overall mean average						*2.11*	*4.41*	*3.83*				

TABLE 10 Feedback on the training sessions, provided by 11 respondents in total.

From 1 to 10, please rate:	*N*	Min.	Max.	Mean	Std. deviation
Rate your own knowledge of cognitive biases before the training	11	2	7	3.90	1.524
Rate your own knowledge of cognitive biases after the training	11	6	9	7.30	0.949
Rate your own knowledge of effective group working (e.g. the Delphi method) before the training	11	1	5	2.20	1.317
Rate your own knowledge of effective group working (e.g. the Delphi method) after the training	11	4	8	6.20	1.229
Rate the applicability of the training for social work practice with children and families	11	6	10	8.30	1.703
Rate how likely you are to recommend this training.	11	7	10	8.90	1.287
Valid *N* (listwise)	11				

TABLE 11 Brier scores for the Clarke case vignette (pre- and post-training).

	N	Min.	Max.	Mean	Std. deviation
Stage 1 survey	33	0.06	1.09	0.65	0.286
Post-training (group aggregate, mean average)	10	n/a	n/a	0.30	n/a
Post-training (group aggregate, extremizing)	10	n/a	n/a	0.22	n/a
Post-training (group aggregate, best-performer)	4	n/a	n/a	0.22	n/a

TABLE 12 Brier scores for the poppy case vignette (pre- and post-training).

	N	Min.	Max.	Mean	Std. deviation
Stage 1 survey	35	0.11	0.83	0.52	0.103
Post-training (group aggregate, mean average)	11	n/a	n/a	0.40	n/a
Post-training (group aggregate, extremizing)	11	n/a	n/a	0.46	n/a
Post-training (group aggregate, best-performer)	3	n/a	n/a	0.41	n/a

TABLE 13 Brier scores for the Michael case vignette (post-training).

	N	Min.	Max.	Mean	Std. deviation
Michael's allocated social worker (Wilkins & Meindl, 2022)	1	n/a	n/a	0.42	n/a
Post-training (group aggregate, mean average)	11	n/a	n/a	0.40	n/a
Post-training (group aggregate, extremizing)	11	n/a	n/a	0.46	n/a
Post-training (group aggregate, best-performer)	3	n/a	n/a	0.41	n/a

whether such interventions can—or should—be adapted for social work, if so how, and to what benefit? Of the 130 people invited to take part, 52% ($n = 68$) completed the survey sufficiently for inclusion. Of these, 32% ($n = 22$) were invited to take part in the stage 2 and 3 training. Of these, 81% ($n = 18$) attended at least one session, and 54% ($n = 12$) attended all three. Finally, 50% ($n = 9$) of the subsample attended at least one stage 4 follow-up session. These figures indicate that, despite their high workloads, a number of social workers and other members of staff were willing to attend training sessions and provide post-training data to help develop and measure the accuracy of their judgements. The same figures also illustrate the challenge of attrition, as the number of respondents fell at each stage, such that by stage 4, only 13.2% ($n = 9$) of those who completed the stage 1 survey were still involved (albeit the majority of these were not invited to participate as per the study design). In combination with the generally positive training feedback (Table 10), these figures suggest that Tetlock et al.'s methods can be used with social workers.

Of course, the judgements and decisions made by social workers are much more complex than being simply a question of accuracy. As such, it is important to ask whether and to what extent increased accuracy would make a difference for social work, and what opportunity-cost there might be relative to other important areas for development, such as supervision more generally (Beddoe & Wilkins, 2019; Wilkins & Jones, 2018). Even if we did agree that increased accuracy is a goal worth seeking, there are various ways it may be achieved—including through the use of structured decision-making tools (Shlonsky & Wagner, 2005) and actuarial risk assessment models (Johnson, 2011). In any case, arguments about the value of accuracy are underpinned by the theory of judgemental rationalism (Bhaskar, 2013) and the claim that veracity is a relational property of reality (Moore, 1901; Russell, 1984). Individuals *can* make rational judgements about the world, albeit there are always limitations to our knowledge. The most we can often hope for is practical adequacy (Sayer, 2010)—to use our knowledge to generate expectations about the world and the likely consequences of our actions. The accuracy of these expectations matters 'because people act upon their beliefs—whether [they] are true or not' (Boghossian & Lindsay, 2019, p. 5). When we make (more) accurate judgements, we can make (more) effective decisions by identifying actions that (i) are in the best interests of the child and family, (ii) are more likely to achieve their intended outcomes and (iii) are more aligned with our goals and values (Hastie & Dawes, 2009; Paternoster & Pogarsky, 2009). This is especially important in high-risk situations where the consequences of these decisions will be most significant (Benbenishty et al., 2015; Healy et al., 2009). Thus, while the judgements made by social workers must always involve moral and practical considerations (Taylor & White, 2006), they also need to be well calibrated (Keren, 1991) and based on sufficiently accurate world-models such that (more) effective decisions can be made, and actions taken that are (more) likely to achieve their intended outcomes (Munro, 2019a).

5.1 | Adapting the interventions and next steps

Based on the learning from this study, there are (at least) two adaptations that could help make these interventions more suitable for social work. First, selecting respondents with above-average existing abilities is useful when your aim is to win forecasting tournaments (Tetlock et al., 2014). It is less useful if your aim is to improve the overall quality of judgement (and decision-making) within diverse social work teams. In future studies, it would be constructive to include a wider sample of respondents, with a range of abilities (of critical reflection, open-minded thinking and pattern recognition), to understand how the same training for cognitive debiasing and group work might improve—or not—the judgemental accuracy of more mixed groups. It would also be beneficial to explore the use of the Delphi method with mixed groups too, for example, comparing groups composed solely of high-performers, and groups with a mixture of high-performers and others.

Second, it is important to think how best to identify questions that really matter for social workers (and for children and families). This study, and others (Wilkins & Meindl, 2022), have shown it is

possible for social workers to answer retrospective questions about case vignettes, and to measure the accuracy of their responses, for example, whether a child will come into care, whether there will be another referral or another police call-out to the home. And yet such questions may be of limited value for ongoing casework. Future studies should focus much more on the issue of what questions to ask, how they might be generated (Gruetzemacher, 2022) and how forming more accurate judgements about such questions could inform social work decision-making. For example, it may be that the questions that matter most to social workers are essentially 'unanswerable' (or at least *endlessly debateable*), such as 'will this child be safe?' or 'will this child's wellbeing be promoted more effectively if they come into care or remain with their family'? The method of reciprocal forecasting (Karger et al., 2021; Karger et al., 2022) whereby people are asked to make forecasts about each other's forecasts may be worth exploring with social workers.

Following this proof-of-concept study, it would be instructive to run the study again with a larger sample, so that, for example, everyone who completes the stage 1 survey is invited to attend training. It would also be worthwhile constructing different types of groups following the training, so that some are composed entirely of high performers, while some are mixed. This would help demonstrate whether and to what extent the efficacy of the training depends on the involvement of high-performing social workers or whether it can be useful for those with less developed skill-sets to begin with. It would also be useful to explore how the same methods work prospectively rather than retrospectively, for example, if social workers were to consider questions in relation to their own ongoing casework (where the outcome is unknown because it has yet to happen) instead of case vignettes (where the outcome has already happened but is unknown to those taking part in the study).

6 | CONCLUSION

The quality of judgements and decisions made by social workers is a central component of humane and effective practice and an important area for research. In this study, a set of interventions developed for use in other fields were tested with social workers for the first time. Ultimately, the judgements and decisions made by social workers are moral and political, as much as they are technical and rational (Taylor & White, 2006). Yet the ability to make more accurate judgements has the potential to help us achieve our moral and political objectives, even while providing a technical-rational way of conceptualizing good judgement (Hammond, 2000). In conclusion, these findings indicate that further, larger and more rigorous studies of these interventions can be warranted for social work.

ACKNOWLEDGEMENTS

Thanks to everyone who took part in the study and to the local authority senior managers who gave permission for the study to go ahead within their organization. Many thanks as well to Ms Melissa Meindl, who has contributed significantly to this programme of work.

CONFLICT OF INTEREST STATEMENT

The author has no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

David Wilkins https://orcid.org/0000-0003-2780-0385

REFERENCES

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

Beddoe, L., & Wilkins, D. (2019). Does the consensus about the value of supervision in social work stifle research and innovation? *Aotearoa New Zealand Social Work*, *31*(3), 1–6. https://doi.org/10.11157/anzswj-vol31iss3id643

Benbenishty, R., Davidson-Arad, B., López, M., Devaney, J., Spratt, T., Koopmans, C., Knorth, E. J., Witteman, C. L. M., del Valle, J. F., & Hayes, D. (2015). Decision making in child protection: An international comparative study on maltreatment substantiation, risk assessment and interventions recommendations, and the role of professionals' child welfare attitudes. *Child Abuse & Neglect*, *49*, 63–75. https://doi.org/10.1016/j.chiabu.2015.03.015

Bhaskar, R. (2013). *A realist theory of science*. Routledge. https://doi.org/10.4324/9780203090732

Boghossian, P., & Lindsay, J. (2019). *How to have impossible conversations: A very practical guide*. Da Capo Lifelong Books.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Butler, I. (2002). A code of ethics for social work and social care research. *British Journal of Social Work*, *32*(2), 239–248. https://doi.org/10.1093/bjsw/32.2.239

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509–526. https://doi.org/10.1017/S1930297500004599

Cross, T. P., & Casanueva, C. (2009). Caseworker judgments and substantiation. *Child Maltreatment*, *14*(1), 38–52. https://doi.org/10.1177/1077559508318400

Cyr, C., Dubois-Comtois, K., Paquette, D., Lopez, L., & Bigras, M. (2022). An attachment-based parental capacity assessment to orient decision-making in child protection cases: A randomized control trial. *Child Maltreatment*, *27*(1), 66–77. https://doi.org/10.1177/1077559520967995

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? The surprising informational value of "just asking". *Judgment and Decision Making*, *14*(2), 135–147. https://doi.org/10.1017/S1930297500003375

de Winter, J. C., Dodou, D., & Eisma, Y. B. (2023). Responses to Raven matrices: Governed by visual complexity and centrality. *Perception*, *52*(9), 645–661. https://doi.org/10.1177/03010066231178149

DePanfilis, D., & Girvin, H. (2005). Investigating child maltreatment in out-of-home care: Barriers to effective decision-making. *Children and Youth Services Review*, *27*(4), 353–374. https://doi.org/10.1016/j.childyouth.2004.11.010

Dickens, J., Howell, D., Thoburn, J., & Schofield, G. (2007). Children starting to be looked after by local authorities in England: An analysis of inter-authority variation and case-centred decision making. *British Journal of Social Work*, *37*(4), 597–617. https://doi.org/10.1093/bjsw/bch276

Farmer, E., & Lutman, E. (2014). Working effectively with neglected children and their families–what needs to change? *Child Abuse Review*, *23*(4), 262–273. https://doi.org/10.1002/car.2330

Featherston, R. J., Shlonsky, A., Lewis, C., Luong, M.-L., Downie, L. E., Vogel, A. P., Granger, C., Hamilton, B., & Galvin, K. (2019). Interventions to mitigate bias in social work decision-making: A systematic review. *Research on Social Work Practice*, *29*(7), 741–752. https://doi.org/10.1177/1049731518819160

Forrester, D. (2008). Child protection and re-referrals involving serious concerns: A follow-up study of 400 referrals closed by social services departments. *Child & Family Social Work*, *13*(3), 286–299. https://doi.org/10.1111/j.1365-2206.2008.00548.x

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, *62*(2), 410–422. https://doi.org/10.1093/isq/sqx078

Gillingham, P., & Humphreys, C. (2010). Child protection practitioners and decision-making tools: Observations and reflections from the front line. *British Journal of Social Work*, *40*(8), 2598–2616. https://doi.org/10.1093/bjsw/bcp155

Gruetzemacher, R. (2022). Bayesian networks vs. conditional trees for creating questions for forecasting tournaments. Available at: https://abnms.org/uai2022-apps-workshop/papers/S5.pdf

Hammond, K. R. (2000). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press on Demand.

Hanea, A. M., McBride, M. F., Burgman, M. A., & Wintle, B. C. (2018). The value of performance weights and discussion in aggregated expert judgments. *Risk Analysis*, *38*(9), 1781–1794. https://doi.org/10.1111/risa.12992

Hastie, R., & Dawes, R. M. (2009). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Healy, K., Meagher, G., & Cullin, J. (2009). Retaining novices to become expert child protection practitioners: Creating career pathways in direct practice. *British Journal of Social Work*, *39*(2), 299–317. https://doi.org/10.1093/bjsw/bcm125

Hilder, J., & Whittaker, A. (2023). Studying the effectiveness of interventions to improve decision making and work with risk. In B. J. Taylor, J. D. Fluke, J. C. Graham, E. Keddell, C. Killick, A. Shlonsky, & A. Whittaker (Eds.), *The SAGE handbook of decision making, assessment and risk in social work* (pp. 569–577). Sage.

Hood, R., Abbott, S., Coughlan, B., Nilsson, D., Duschinsky, R., Parker, P., & Mannes, J. (2022). Improving the quality of decision making and risk assessment in children's social care: A rapid evidence review. Available at: https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_Improving_Decision_Making_rapid_review_April2022.pdf

Janssen, E. M., Verkoeijen, P. P., Heijltjes, A. E., Mainhard, T., van Peppen, L. M., & van Gog, T. (2020). Psychometric properties of the actively open-minded thinking scale. *Thinking Skills and Creativity*, *36*, 100659. https://doi.org/10.1016/j.tsc.2020.100659

Johnson, W. L. (2011). The validity and utility of the California family risk assessment under practice conditions in the field: A prospective study. *Child Abuse & Neglect*, *35*(1), 18–28. https://doi.org/10.1016/j.chiabu.2010.08.002

Karger, E., Atanasov, P. D., & Tetlock, P. (2022). Improving judgments of existential risk: Better forecasts, questions, explanations, policies. *Questions, Explanations, Policies (January 5, 2022)*.

Karger, E., Monrad, J., Mellers, B., & Tetlock, P. (2021). Reciprocal scoring: A method for forecasting unanswerable questions. Available at SSRN 3954498.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*, 217–273. https://doi.org/10.1016/0001-6918(91)90036-Y

Kirkman, E., & Melrose, K. (2014). *Clinical judgement and decision-making in children's social work: An analysis of the 'front door' system*. London: Department for Education.

Lietz, C. A. (2008). Implementation of group supervision in child welfare: Findings from Arizona's supervision circle project. *Child Welfare*, *87*(6), 31–48.

Lietz, C. A. (2009). Critical thinking in child welfare supervision. *Administration in Social Work*, *34*(1), 68–78. https://doi.org/10.1080/03643100903432966

Linstone, H. A., & Turoff, M. (1975). *The Delphi method*. Addison-Wesley Reading.

Masson, J. (2006). The climbie inquiry - context and critique. *Journal of Law and Society*, *33*(2), 221–243.

Mathias, J. (2015). Thinking like a social worker: Examining the meaning of critical thinking in social work. *Journal of Social Work Education*, *51*(3), 457–474. https://doi.org/10.1080/10437797.2015.1043196

McLeod, J. W., & McCrimmon, A. W. (2021). *Test review: Raven's 2 progressive matrices* (Clinical ed.). SAGE Publications.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., Tetlock, P., & Arkes, H. R. (2019). Forecasting tournaments, epistemic humility and attitude depolarization. *Cognition*, *188*, 19–26. https://doi.org/10.1016/j.cognition.2018.10.021

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making*, *12*(4), 369–381. https://doi.org/10.1017/S1930297500006240

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565. https://doi.org/10.1287/mnsc.2016.2525

Moore, G. (1901). Truth and falsity. In *Selected writings*. Routledge.

Munro, E. (2019a). Decision-making under uncertainty in child protection: Creating a just and learning culture. *Child & Family Social Work*, *24*(1), 123–130. https://doi.org/10.1111/cfs.12589

Munro, E. (2019b). *Effective child protection*. SAGE Publications Limited.

Paternoster, R., & Pogarsky, G. (2009). Rational choice, agency and thoughtfully reflective decision making: The short and long-term consequences of making good choices. *Journal of Quantitative Criminology*, *25*, 103–127. https://doi.org/10.1007/s10940-009-9065-y

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Raven, J. (2008). The Raven progressive matrices tests: Their theoretical basis and measurement model. In *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics* (pp. 17–68). Royal Fireworks Press.

Russell, B. (1984). *Theory of knowledge: The 1913 manuscript* (Vol. 7). Routledge.

Sayer, A. (2010). *Method in social science: Revised* (2nd ed.). Routledge.

Schmidt, B. (2006). Proof of principle studies. *Epilepsy Research*, *68*(1), 48–52. https://doi.org/10.1016/j.eplepsyres.2005.09.019

Scottish Executive. (2006). Report of the 21st-century social work review. Edinburgh: Scottish Executive.

Shlonsky, A., & Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management. *Children and Youth Services Review*, *27*(4), 409–427. https://doi.org/10.1016/j.childyouth.2004.11.007

Stanovich, K. E., & Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence*, *11*(2), 27. https://doi.org/10.3390/jintelligence11020027

Taylor, B. (2017). *Decision making, assessment and risk in social work*. Learning Matters.

Taylor, B., & Whittaker, A. (2018). Professional judgement and decision-making in social work. *Journal of Social Work Practice*, *32*, 105–109.

Taylor, C., & White, S. (2006). Knowledge and reasoning in social work: Educating for humane judgement. *British Journal of Social Work*, *36*(6), 937–954. https://doi.org/10.1093/bjsw/bch365

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, *23*(4), 290–295. https://doi.org/10.1177/0963721414534257

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Wilkins, D., & Jones, R. (2018). Simulating supervision: How do managers respond to a crisis? *European Journal of Social Work*, *21*(3), 454–466. https://doi.org/10.1080/13691457.2017.1366429

Wilkins, D., & Meindl, M. (2022). Can child protection social workers forecast future actions, events and outcomes? A case study of long-term work with five families. *Child Care in Practice*, 1–20. https://doi.org/10.1080/13575279.2022.2118674

How to cite this article: Wilkins, D. (2024). Improving the accuracy of social work judgements: A proof-of-concept study of a training programme. *Child & Family Social Work*, *29*(4), 948–959. https://doi.org/10.1111/cfs.13146