

A randomised controlled trial of a 'checklist' intervention to mitigate confirmation bias and improve forecasting accuracy in social work

Melissa Meindl * and David Wilkins 

Children's Social Care Research and Development Centre (CASCADE), School of Social and Spatial Sciences, Cardiff University, sbarc|spark, Maindy Rd, Cardiff, CF24 4HQ, UK

*Corresponding author. Children's Social Care Research and Development Centre (CASCADE), School of Social and Spatial Sciences, Cardiff University, sbarc|spark, Maindy Rd, Cardiff, CF24 4HQ, UK. E-mail: Meindlm@cardiff.ac.uk.

Abstract

Social workers may be particularly at risk of cognitive biases because they make judgments and decisions under uncertain conditions, involving predicting future outcomes. Given the potential consequences, evidence is needed on whether bias mitigation tools improve forecasting accuracy in social work. This study tested the impact of a checklist intervention on social workers' forecasting accuracy (Brier scores) and confirmation bias (Wason Selection Task scores) using vignettes from real referrals. In an online survey, 88 participants responded to baseline vignettes before being randomly allocated to a control or intervention group. The intervention group completed the checklist before responding to endline vignettes and the Wason Selection Task, while the control group accessed the checklist afterwards. The intervention group's mean Brier score decreased (0.44 to 0.41), indicating improved forecasting accuracy, while the control group's increased (0.44 to 0.45), indicating reduced forecasting accuracy. These differences were not statistically significant. Wason Selection Task scores were similar across groups, indicating that the checklist made no difference to confirmation bias. Despite its effectiveness in other fields, a brief online checklist may not be sufficient to improve forecasting accuracy or

© The Author(s) 2026. Published by Oxford University Press on behalf of
The British Association of Social Workers.
This is an Open Access article distributed under the terms of the Creative
Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in
any medium, provided the original work is properly cited.

confirmation bias in social work. Future research should explore practice-embedded or supervisory approaches in real practice settings rather than online.

Keywords: Brier scores; checklist intervention; cognitive bias mitigation; forecasting accuracy; social work decision-making.

Accepted: February 2026

Introduction

When people make judgements and decisions, particularly if seeking solutions to problems, their brains may rely on cognitive shortcuts, also called heuristics. These can be understood as the brain's attempt to simplify information processing by relying on 'rules of thumb' or things that worked well in the past. Heuristics can be useful when making decisions quickly; however, they can also result in cognitive biases, which are systematic errors in thinking and judgement (Tversky and Kahneman 1974; Elstein 1999; Gilovich et al., 2002; Gigerenzer and Gaissmaier 2011).

There are more than 100 recognised cognitive biases, many of which are known to affect judgement and decision-making in a wide range of areas (Heick, n.d.). For example, anchoring bias describes how we may rely more heavily on the first piece of information we have, interpreting subsequent information in relation to the 'anchor' reference point (Tversky and Kahneman 1974). This may happen in social work when a referral indicates that a child is at significant risk of harm; new information about the family may be interpreted only in light of this initial judgement, rather than more objectively.

It is not currently known which cognitive biases are most relevant, prevalent, and possibly detrimental for decision-making in social work. Consequently, there is a lack of research and evidence on cognitive bias mitigation within the social work domain (Featherston et al., 2019). This is problematic because social workers make decisions that can have significant consequences for children and families, for example, whether to accept a referral, undertake a child protection investigation, pursue care proceedings, or end their involvement with a family. Many of these decisions involve implicit or explicit predictions or judgements about the likelihood of different future outcomes.

English (and Welsh) law requires social workers to make judgements about the likelihood of future significant harm. Such judgements involve making '*a prediction from existing facts, often from a multitude of such facts, about what has happened in the past, about the characters and personalities of the people involved, [and] about the things which they have said and done*' (House of Lords 2008). More generally, social workers are making constant

judgements about a range of issues, such as whether a child's needs will be better met within a kinship care or foster care placement, parental capacity for change, and the impact of different interventions.

However, predicting the future is rarely easy, and social workers are particularly vulnerable to cognitive error because the decisions they make are about human behaviour and informed by often incomplete and uncertain information (Munro 2011).

Confirmation bias

Confirmation bias is a type of cognitive bias where individuals seek out, interpret, and recall information in ways that confirm their existing beliefs, expectations, or values, while discounting contradictory evidence (Nickerson 1998). In professional decision-making, this bias can contribute to systematic errors in judgement (Kahneman 2011). In a social work context, confirmation bias may, for example, lead social workers to pay attention only to information that supports their initial thoughts about a family, or confirms the result of a risk assessment, while ignoring or undervaluing evidence that challenges those views (O'Sullivan 2011).

Confirmation bias was targeted for intervention in this study for two key reasons: (1) it was one of the three most prevalent biases identified by social workers as impacting social work decision-making in previous research (see the 'Previous research' section below), and (2) due to the availability of relevant debiasing tools and standardised measures of confirmation bias compared to the other identified cognitive biases.

Previous research

In a previous randomised controlled trial (RCT), Wilkins et al. (2020) tested three brief online interventions: (i) confidence calibration, (ii) cognitive debiasing, and (iii) growth mindset-based feedback, to improve social workers' probabilistic judgement accuracy. Participants were asked to read four vignettes based on real, anonymised referrals to children's services. After reading each vignette, participants answered questions about the likelihood of different outcomes. Between the second and third vignettes, participants were randomly assigned to one of the three interventions or a control condition. Judgement accuracy was measured via the correspondence between participants' forecasts and the known real-world outcomes of each vignette (i.e. forecasting accuracy). Findings from the RCT showed that none of the three interventions significantly improved forecasting accuracy, although the cognitive debiasing intervention appeared to be the most promising. Overall, social workers found it difficult to predict what would happen next in relation to the vignettes, a finding that is consistent

with wider literature showing that cognitive biases can adversely affect judgement and decision-making across a range of professional domains (Tversky and Kahneman 1974; Murata *et al.*, 2015; Saposnik *et al.*, 2016; Berthet 2022).

In a follow-up study, Meindl and Wilkins (2025) found that social workers' forecasting accuracy was reasonably stable over time, with some variability associated with the nature of the vignettes about which forecasts were made. These findings mirror evidence of variability in social work decision-making more broadly (Keddell and Hyslop 2019). They also suggest that if interventions can be identified that mitigate the influence of cognitive biases and lead to more accurate forecasts, such improvements may be sustainable over at least a medium-term timeframe.

The same follow-up study explored the perceived prevalence of cognitive biases in child and family social work. Social workers ($n=63$) were provided with brief definitions and practice-based examples of 10 cognitive biases and asked how often they believed each bias influenced social work decision-making. All 10 cognitive biases were perceived to have some impact on social work decision-making. The three biases social workers identified as impacting social work decision-making frequently or very frequently were authority bias (64%), blind-spot bias (57%) and confirmation bias (49%), lending further weight to the case for identifying, developing, and testing bias-mitigation interventions in social work.

The present study builds on this programme of research on social worker judgement and decision-making by testing a more in-depth intervention and measuring its effects on social workers' forecasting accuracy and cognitive bias (for the full details of the present study and project report from which this paper stems, see Wilkins and Meindl 2021). Notably, wider research on judgement and decision-making suggests that forecasting accuracy is shaped not only by discrete interventions, but also by factors such as professional experience, individual differences, and features of the decision-making environment itself (e.g. Munro 2011; Baumann *et al.*, 2013; Tetlock and Gardner 2016). These considerations are therefore relevant for interpreting intervention effects and for understanding the potential limits of individual-level debiasing tools.

Piloting potential interventions

To determine which intervention would be tested in this study, relevant literature on cognitive debiasing was reviewed to identify potential interventions that could help mitigate the effect of cognitive biases. Nine interventions were identified. Using two main criteria, (i) existing evidence of effect in other fields and (ii) relevance to the cognitive biases identified in the follow-up study as most impacting social work decision-making, four interventions were chosen for piloting with social work experts recruited

via the Frontline charity. The purpose of piloting was to obtain qualitative feedback on which intervention was appropriate for, and adaptable to, the social work context. The participants ($n = 9$) were 25–54 years of age, and all held a bachelor's degree or higher. Seven were female, and seven were White British. Based on the feedback from this pilot, a 'checklist' intervention was chosen for the present study.

Checklists

Checklists have been widely used in settings such as healthcare to improve decision-making (Gawande 2010) and as tools to address cognitive biases across a range of other domains (Ely et al., 2011; Emmons et al., 2018; Draws et al., 2021). There is no single checklist suitable for universal application; rather, checklists typically consist of a series of prompts developed in a bespoke way, depending on the context, the people involved, and the types of decisions being made. There is also no fixed format for what a checklist should include, although some broad principles are commonly emphasised: for example, checklists should be as short as possible, use simple language, and prompt reflection rather than replace professional judgement (Gawande 2010).

The primary aim of a checklist is to ensure that key components of the judgement and decision-making process are not overlooked. Checklists are not a simple solution, and while they have been shown to improve outcomes in some forms of decision-making, they do not provide a universal solution (Catchpole and Russ 2015).

In other applied decision-making domains, including healthcare, aviation, and risk management, well-designed checklists have been shown to reduce error, improve consistency, and support reflective revision of initial judgements, particularly under conditions of uncertainty. On this basis, it was reasonable to hypothesise that a similar intervention might improve forecasting accuracy in social work. However, there is an absence of prior empirical tests of such interventions in this domain. Existing systematic reviews and the broader social work literature indicate limited evidence for bias-mitigation tools designed specifically for child and family social work (Featherston et al., 2019). Given well-documented variability in child-protection decision-making (Keddell 2014), testing a pragmatic checklist intervention in this context therefore fills an important gap in the research literature.

Research questions

There were two primary research questions for this study:

1. What is the impact of the checklist intervention on forecasting accuracy among social workers?

- 2. What is the impact of the checklist intervention on confirmation bias among social workers?

There were also two additional exploratory research questions:

- 3. Is there a relationship between social workers' forecasting accuracy and their level of confirmation bias?
- 4. Is there a relationship between social workers' forecasting accuracy and (i) age-group, (ii) gender, and (iii) length of post-qualifying experience?

Materials and methods

Study design

This study was an RCT with two arms: one intervention group and one control group, and two outcome measures: one pre- and post-test, and one cross-over (Fig. 1).

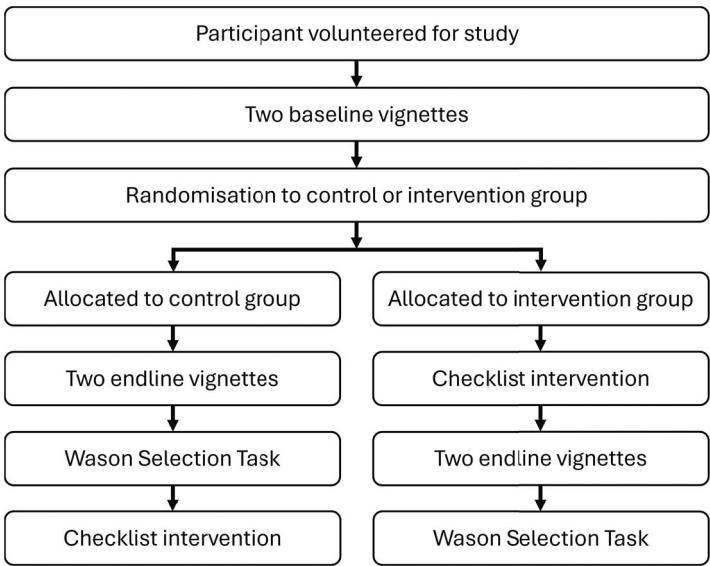


Figure 1. Diagram of cross-over design.

The full trial protocol for the study was preregistered with the Open Science Framework on 4 December 2020 (<https://osf.io/57b84>), prior to data collection. This research project was reviewed and approved by the Cardiff University School of Social Sciences Research Ethics Committee (reference: 3880).

Materials, instruments, and measures

Checklist intervention

The checklist intervention used in this study was originally intended to be tested in-person; however, it had to be delivered online via a survey due to the COVID-19 pandemic. To the researchers’ knowledge, checklists have not been implemented in this way before in social work, and therefore the checklist was piloted with a small group of social workers and social work academics (n=9). Their feedback was used to make minor changes to the checklist and to finalise the items included in it.

The final version of the checklist intervention involved reading a vignette and making three recommendations: two regarding what *should happen* next, and one about what *should not happen* (see Box 1 for an overview). Participants were then asked to read through a checklist (see Table 1) to help them reflect on their initial recommendations and were given the opportunity to revise them.

Box 1: An overview of the vignette used as part of the checklist intervention

Two White British children are referred—Sophie, aged 9, and Madeline, aged 3. Madeline was brought to hospital by her parents with burns to her neck, back, and arms. The father said it was an accident but was also overheard telling the children not to talk to anyone about what really happened. You, the social worker, have been allocated to complete an assessment. When you visit the home, you find the family are struggling financially and observe that the mother and children seem wary of the father. You also discover that the father has another child from a previous relationship, who lives in another part of the country with her mother. You check with the relevant local authority and find that this child used to live with her father and had a history of apparently accidental injuries. The next day, you come into work to find a report saying that police were called to the home last night because of concerns about domestic violence.

*Concluding your assessment, what **two** recommendations would you make about what should happen now? What **one** thing would you recommend should not happen?*

Table 1. The checklist intervention used in this study.

Item	Yes	No	Unsure
1. Have you identified two different options that you would recommend ?			
2. Have you identified one option you would not recommend ?			
3. Can you identify at least one pro and one con for all three recommendations?			
4. Have you actively sought out information in the case study that might contradict your recommendations?			
5. Can you see how another social worker might come up with a different set of recommendations?			
6. Is it possible that your analysis of the case study is being influenced by cognitive biases such as confirmation bias?			
7. Is it possible that you feel too confident about your recommendations?			
8. Would you recommend the same options if the family were of a different ethnicity or social class?			

Vignettes

All vignettes used in the study were selected from a sample of real referrals to children's services provided by an English local authority. Information in the relevant files regarding what happened following each referral was used to generate the questions about each vignette and then to evaluate the accuracy of participants' forecasts, as the outcomes were known to researchers. For example, for the baseline and endline vignettes, participants were asked how likely it was that the child would come into care within the following 12 weeks, using a scale from 0 to 100, where 0 indicates *definitely will not happen* and 100 indicates *definitely will happen* (see [Box 2](#)). This provided data that could be used to determine the extent to which participants were able to make accurate predictions about what

Box 2: An example of a vignette

Referral received from the police. Prior to this referral, the family were not known to children's services.

'Salma (aged 15) has sent naked full-frontal pictures showing her face via Snapchat to an older male she knows as Amal. Salma's parents discovered the images of their daughter on her mother's phone. Due to their grave concerns, they reported the matter to the police. Salma states she has been talking to Amal for around 2 weeks and that she sent the images of herself willingly. When he requested more photos, she refused, and he blocked her. Internet safety has been discussed with parents, they acted appropriately in reporting the matter to the police. From the information available, there appears to be a risk of sexual exploitation and distribution of indecent images of a child'.

Q1. In response to this referral, how likely is each of the following outcomes: (forecasts 1-4)

- No further action
- Social work or other form of assessment
- Emergency removal into care
- Something else

Q2. Within the next 12 weeks, how likely is it that Salma will become the subject of: (*forecasts 5-8*)

- No plan
- Child in need plan
- Child protection plan
- Looked after child plan

Q3. Within the next 12weeks, will the authority convene a strategy meeting? (forecast 9)

Q4. Within the next 6 months, will there be a further referral about this child from any source?
(forecast 10)

Definitely will not happen	Highly unlikely but not impossible	Very Unlikely	Unlikely	Probably won't happen	May or may not happen	Could happen	Quite likely	Very Likely	Highly likely but not definite	Definitely will happen
0	10	20	30	40	50	60	70	80	90	100

Note: The italicised text explains how 10 forecasts resulted from four questions and was not presented to participants.

happened following the referral. Explanatory text in the survey for the baseline and endline vignettes asked respondents to forecast the likelihood of different possible outcomes, rather than to think about what they might have done with a similar referral. Participants were not asked to make their own decision or recommendation about what *should* have happened following the referral, nor to make any judgements about the real-life decisions that had been made. In simple terms, the vignettes (baseline and endline) and associated questions allowed the accuracy of participants' forecasting abilities to be tested, not the inter-rater reliability of decision-making in social work.

Outcome measures

Brier scores (Brier 1950) were used to measure forecasting accuracy. They are the sum of the squared differences between predicted probabilities and actual outcomes. As described above, participants made 10 forecasts per vignette (each forecast ranging from 0% to 100%). Forecasts were elicited independently for each outcome and, therefore, were not required to sum to 100% across outcome options within the same question. This format was used because the aim was to capture perceived likelihoods rather than enforce probabilistic coherence across mutually exclusive options.

Forecast percentages were converted to proportions for analysis (e.g. 40% → 0.40), each outcome option was coded as a binary event (1 if the outcome occurred, 0 if the outcome did not occur), and then Brier scores were calculated using the following formula: $(1-x)^2 + (0-y)^2 = z$, where x is the participant's forecasted probability that the outcome option occurs, y is the participant's forecasted probability that the outcome does not occur (see Box 3 for a worked example), and z is the Brier score. The 10 binary Brier scores for each vignette were averaged to produce a vignette-level Brier score. Vignette-level Brier scores for the two baseline vignettes were averaged to determine participants' pre-intervention Brier score. Vignette-level Brier scores for the two endline vignettes were averaged to determine participants' endline Brier score. Participants' Brier scores for all four vignettes were averaged to determine their overall Brier score. Brier scores ranged between 0 and 2. Lower Brier scores indicate greater forecasting accuracy, with a score of 0 representing perfect accuracy and a score of 2 representing maximal inaccuracy. Flipping enough coins (equivalent to guessing in relation to questions with a binary outcome) would result in a mean Brier score of 0.5.

The Wason Selection Task (Wason 1968), also called the four-card problem, was used to measure confirmation bias. It is a logic puzzle designed by British psychologist Peter Cathcart Wason to test participants' ability to identify disconfirming information in response to a series of questions. It consists of four cards with a conditional rule, such as '*if a card has a circle on one side, then it has the colour yellow on the other side*'. Only one side

Box 3: A worked example of calculating Brier scores
Consider Q1 in the example vignette provided in [Box 2](#). Suppose a participant provides the following forecasts for how likely they think each outcome is:

- No further action: 40%
- Social work or other form of assessment: 35%
- Emergency removal into care: 15%
- Something else: 10%

Assume the actual observed outcome was 'Social work or other form of assessment'.

Step 1: Convert % to proportions: 40% → 0.40, 35% → 0.35, 15% → 0.15, 10% → 0.10

Step 2: Calculate the Brier score (BS) for each outcome

- No further action (outcome did not occur; 0):

$$BS = (0 - 0.40)^2 + (1 - 0.60)^2 = 0.32$$

- Social work or other form of assessment (outcome occurred; 1):

$$BS = (1 - 0.35)^2 + (0 - 0.65)^2 = 0.845$$

- Emergency removal into care (outcome did not occur; 0):

$$BS = (0 - 0.15)^2 + (1 - 0.85)^2 = 0.045$$

- Something else (outcome did not occur; 0):

$$BS = (0 - 0.10)^2 + (1 - 0.90)^2 = 0.02$$

Step 3: Repeat Brier score calculation for each forecast item within the vignette and then average all 10 to get participant's vignette-level Brier score.

of each of the cards is shown, and participants are asked to identify the card(s) that need to be turned over to determine whether the rule is valid.

The task tests confirmation bias by measuring whether participants turn over cards that could confirm or disconfirm the rule. Participants were awarded one point for each card turned over correctly (to disconfirm the rule) and one minus point for each card turned over incorrectly (to confirm the rule). As participants in this study were presented with three card sets, the highest possible score was +6 (indicating low levels of confirmation bias) and the lowest possible score was -6 (indicating high levels of confirmation bias).

Participants

Recruitment

Participants were recruited using advertisements on the What Works for Social Care website (now Foundations: www-foundations.org.uk) and social media platforms. Partner local authorities, other social work organisations in England (e.g. Frontline), Cardiff University, and the Children's Social

Care Research and Development Centre (CASCADE) also publicised the study through their respective newsletters and social media platforms.

Participants were required to be a registered social worker or student social worker on social work qualifying programmes in England. Participants were asked to self-certify that they met this criterion; any participants who declined to do so were not eligible for participation.

Participants who completed the study were offered a £10 gift voucher as a ‘thank you’ for taking part. All data were collected anonymously unless participants opted to provide their email address for the purpose of sending the gift vouchers and/or to be contacted for future studies. Where participants chose to provide their email address, it was not linked with any data they provided.

Sample demographics

Eighty-eight participants were included in the full analysis. Most were female, and the single largest age group was 25–34 years old. More than a third of participants had been practicing social work for 10 years or more (Table 2). As participants signed up voluntarily and without specific targeting of particular groups, the sample is not representative of the wider social work population in England.

Table 2. Participants’ personal and professional characteristics (n = 88).

Characteristic		N	%
Gender	Male	14	15.9
	Female	73	83.0
	Non-binary	1	1.1
Age	18–24	5	5.7
	25–34	35	39.8
	35–44	25	28.4
	45–54	11	12.5
	55–64	11	12.5
	65–74	1	1.1
Length of experience	N/A—I am not yet qualified	1	1.1
	Between 1 and 12 months	10	11.4
	Between 1 and 3 years	19	21.6
	Between 4 and 6 years	14	15.9
	Between 7 and 9 years	10	11.4
	10 or more years	34	38.6

Previous participation

At the outset of the survey, participants were asked to declare whether they had taken part in any of the researchers’ previous studies on social worker

decision-making (Wilkins *et al.*, 2020; Meindl and Wilkins 2025). Seventeen (19.3%) participants said they had. As some of these involved the same or similar vignettes, efforts were made to ensure key details (e.g. names) were changed for this study. However, there was a chance these participants had seen the vignettes before and would be familiar with them. To determine if they performed differently from other participants, an independent samples t-test was used to compare their mean Brier scores with those of participants who had not taken part in previous studies. No statistically significant differences in Brier scores were found for participants with self-reported previous experience of similar studies ($M=0.43$, $SD=0.12$) compared with participants without such experience ($M=0.43$, $SD=0.09$), $t(86)=0.20$, $P=.84$, $d=0.05$ (two tailed). The mean difference in Brier scores between the two groups of participants was 0.01 with a 95% confidence interval ranging from -0.05 to 0.06 . This suggests that participants who took part in previous, similar studies did not perform differently compared with the other participants and were included in the data analysis.

Data collection

The survey was hosted on Qualtrics (<https://www.qualtrics.com>) between December 2020 and February 2021 and took approximately 30–60 minutes to complete.

At the start of the survey, participants were asked questions about their professional and personal characteristics. Participants then read two baseline vignettes and answered four questions after each one, involving making 10 forecasts per vignette regarding the likelihood of different outcomes (see Box 2 for an example of how the vignettes were presented, including the questions posed about different outcomes, and a copy of the scale that participants were asked to use when providing their forecasts).

Following the two baseline vignettes, participants were randomly allocated to the control group or the intervention group. Participants in the intervention group were asked to complete the checklist intervention first, before reading and answering questions about two endline vignettes, and then completing the Wason Selection Task (see the ‘outcome measures’ section). Participants in the control group were asked to complete the two endline vignettes and then the Wason Selection Task, before accessing the checklist intervention (see Fig. 2).

The endline vignettes again involved answering four questions about each one, involving making 10 forecasts per vignette regarding the likelihood of different outcomes. Participants who completed the whole survey were asked to make 40 forecasts in total.

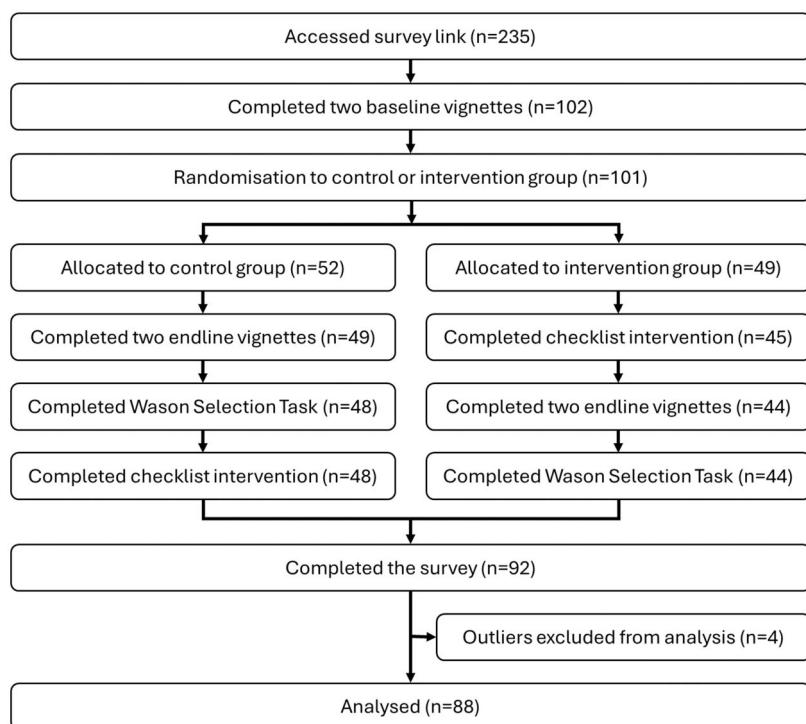


Figure 2. A CONSORT flow diagram overview of the sample throughout the study.

Data analysis

Full details of the analysis are provided in the trial protocol (<https://osf.io/57b84>). Participants and analysts were blind to group allocation. All data cleaning, normality assessments, and analyses were carried out using IBM SPSS Statistics (Statistical Package for the Social Sciences; Version 25) and Microsoft Excel.

Overall, 235 participants accessed the survey link. Those who did not complete $\geq 75\%$ of the survey were excluded from analysis ($n = 143$). Within the remaining 92 participants, 11 outliers were identified using extreme values, QQ plots, and boxplots. Four of the outliers were excluded from analysis as their responses indicated they may not have understood the survey. The remaining outliers were included in data analysis as a comparison of original and 5% trimmed means showed the values did not strongly influence the data. All exclusions were made blind to group allocation.

Kolmogorov–Smirnov values indicated the assumption of normality was not violated. Histograms and Q-Q plots confirmed relatively normal distributions; however, mean Brier scores were slightly positively skewed (clustered toward

lower Brier score values, indicating greater accuracy). This was expected given the sample consisted of social workers/student social workers, and they were being asked to forecast outcomes within their professional domain, about which one can assume they have relevant expertise and experience.

Findings

Checklist intervention effects on forecasting accuracy

Multiple regression was used to determine the impact of the checklist intervention on forecasting accuracy. Endline Brier scores were regressed on condition allocation, baseline Brier scores, age, gender, and length of post-qualifying experience. No assumptions were violated. The full model explained 5.2% of the variance in endline Brier scores (adjusted $R^2 = 0.05$, $F(5, 82) = 1.96$, $P = .09$). The coefficient on the condition allocation dummy was used to determine the treatment effect and indicated that the checklist intervention did not have a significant impact on forecasting accuracy ($\beta = -0.127$, $P = .23$, $SE = 0.03$), with condition allocation explaining only 1.5% of the variance in endline Brier scores. Table 3 and Fig. 3 show a small increase in the Brier score for the par-

Table 3. Baseline and endline Brier score comparison.

Time-point of Brier score	Control (N = 44)		Intervention (N = 44)	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Baseline Brier score	0.44 (0.12)	0.45 (0.17)	0.44 (0.10)	0.45 (0.15)
Endline Brier score	0.45 (0.14)	0.42 (0.21)	0.41 (0.12)	0.41 (0.12)
Baseline-endline difference in Brier score	0.01 (0.02)	-0.03 (0.04)	-0.03 (0.02)	-0.04 (-0.03)

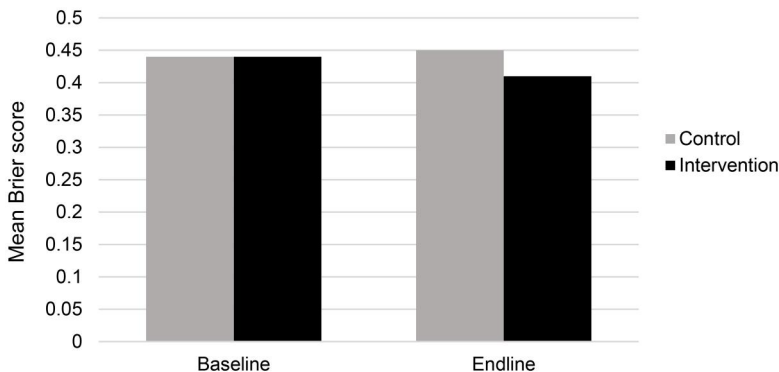


Figure 3. Comparison of baseline and endline mean Brier scores.

ticipants in the control group (increased by 0.01), and a small decrease in the Brier score for the intervention group between the baseline and endline measurements (decreased by 0.03). This indicates that participants' forecasting accuracy improved after the checklist intervention, whereas forecasting accuracy decreased for participants who did not complete the checklist intervention. However, these changes were small, and the difference between the two groups was not statistically significant.

Checklist intervention effects on confirmation bias

Multiple regression was used to determine the impact of the checklist intervention on confirmation bias. Wason Selection Task scores were regressed on condition allocation, age, gender, and length of post-qualifying experience. No assumptions were violated. The full model explained 2.3% of the variance in Wason Selection Task Scores (adjusted $R^2 = 0.02$, $F(4, 82) = 1.50$, $P = .21$). The intervention group achieved a slightly lower score on the Wason Selection Task (1.68) compared with the control group (1.93). A higher score on this measure indicates less confirmation bias. However, the coefficient on the condition allocation dummy was used to determine the treatment effect and indicated that the checklist intervention did not have a significant impact on confirmation bias ($\beta = -0.10$, $P = .38$, $SE = 0.27$), with condition allocation explaining only 1% of the variance in Wason Selection Task scores (Fig. 4). This indicates that the checklist intervention appeared to make no meaningful difference to confirmation bias.

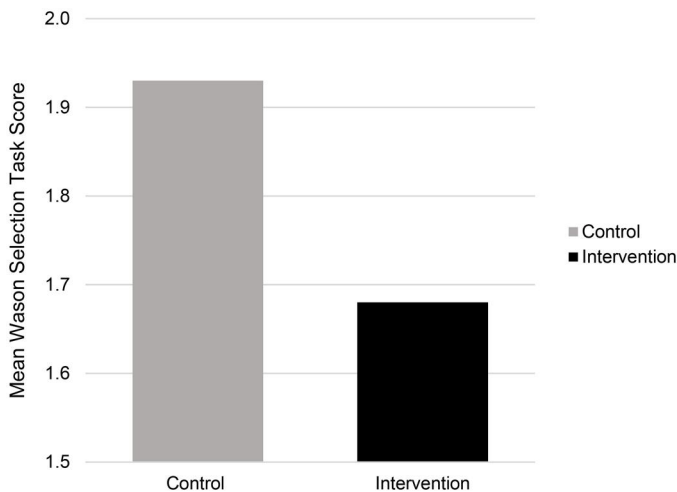


Figure 4. Between-group comparison of Wason Selection Task scores.

Relationship between forecasting accuracy and confirmation bias

Pearson’s product-moment correlation coefficient was used to explore if there was a relationship between social workers’ forecasting accuracy (Brier scores) and their level of confirmation bias (Wason Selection Task scores). There was a very small negative correlation between the Brier scores and Wason Selection Task scores that did not reach statistical significance ($r = -0.10$, $n = 87$, $P = .37$). This indicates there was no significant relationship between participants’ forecasting accuracy and their level of confirmation bias.

Relationship between forecasting accuracy and professional/ personal characteristics

The multiple regression model exploring intervention effects on forecasting accuracy was also used to determine whether there was a relationship between social workers’ forecasting accuracy and personal characteristics: (i) age-group, (ii) gender, and (iii) length of post-qualifying experience.

Baseline Brier scores made the strongest unique contribution to endline Brier scores and were the only significant predictor of endline Brier scores ($\beta = 0.225$, $P = .04$). None of the personal or professional characteristic variables significantly predicted Brier scores (Table 4). This indicates that participants’ personal and professional characteristics were not associated with their forecasting accuracy or level of confirmation bias.

Table 4. Personal and professional characteristics and Brier scores ($n = 88$).

Characteristic	Standardised coefficients (Beta)	Std. error	t value	P value	% of variance explained
Baseline Brier score	0.225	0.129	2.14	.035	4.9%
Gender	−0.146	0.037	−1.38	.172	2%
Age	0.021	0.014	0.17	.865	0%
Length of post-qualifying experience	0.141	0.011	1.13	.260	1.3%

Discussion

It is challenging to make accurate forecasts about the future, particularly within the complex environment of social work. Given the evidence that checklist-based interventions can improve decision-making, reduce cognitive bias, and reduce error in other professional fields, it was reasonable to expect that the checklist tested in this study might improve the forecasting accuracy of social workers. We found that the intervention group’s mean

Brier score reduced from 0.44 to 0.41 (indicating improved forecasting accuracy), while the control group's increased from 0.44 to 0.45 (indicating reduced forecasting accuracy). Although the trend favoured the intervention group, the difference was not statistically significant, indicating that the checklist intervention made no significant difference to forecasting accuracy. The checklist also made no meaningful difference to confirmation bias, with similar Wason Selection Task scores across both groups. This suggests that the mechanisms through which checklists improve decision-making in other fields may not readily transfer to social work forecasting, at least when delivered as a brief, individual, online intervention.

Beyond the primary outcomes, several broader considerations help contextualise these findings. First, what does 'accurate' forecasting look like in social work? It seems reasonable to expect that social workers, all else being equal, should be able to make forecasts more accurately than would be expected by chance. In a previous study, [Meindl and Wilkins \(2025\)](#) found that a group of 283 social workers achieved an average Brier score of 0.47 (6% better than you would expect by chance). Similar results were found in this study, with an average pre-intervention Brier score of 0.44 and an average endline Brier score of 0.43.

Second is the nature of the environment within which social workers are making forecasts. Making accurate forecasts is relatively easy when conditions are stable and predictable. However, the practice environment of child and family social work is changeable and complex. Within this context, there is much variation in outcomes that may be easier or more difficult to forecast. For example, some vignettes used in this study have been consistently easier to forecast accurately compared with others, which may provide an explanation for the variable performance of the participants. Social workers were also being asked to make these forecasts in a simulated setting, whereas in any real-life setting, social workers have access to far more information about the families they are working with that may allow them to make more accurate, contingent forecasts about the likelihood of different outcomes.

Finally, it is important to discuss the proposed relationship between forecasting, judgement, and decision-making in social work. Baumann et al.'s model (2013) shows how the ecology of social work decision-making involves many different factors, and that the role of the individual decision-maker can only be understood alongside the influence of other factors, including external, case-related and organisational. At the individual level, Dalglish's model of assessment and decision-making ([Dalglish 1988](#)) is a useful guide for understanding how various factors influence the formation of a judgement about a situation. Depending on the nature of a judgement, a particular threshold for taking action may or may not be crossed. This helps to understand the difference between making a judgement and making a decision. One of the aims of this study, and the wider programme of research on social worker decision-making that it

stems from, is to explore how accurately social workers make forecasts and what helps promote greater accuracy, to test whether forecasting accuracy is a helpful proxy measure for at least some element of 'good judgement'. If so, Brier scores could offer a useful way of testing the effectiveness of different interventions aimed at improving social work judgement (and decision-making).

Perhaps the most notable finding from this study was the lack of a correlation between personal and professional characteristics and forecasting accuracy. It may be assumed that more experienced social workers would make more accurate forecasts. Indeed, in some studies, less experienced workers have found it harder to analyse information (Drury-Hudson 1999). On the other hand, Tetlock's work has shown that even very experienced experts can struggle to make accurate forecasts, and that some elements of greater experience are negatively correlated with forecasting accuracy (Tetlock and Gardner 2016). For example, increased experience may lead to increased confidence, and perhaps over-confidence at times (Sanchez and Dunning 2018). It is possible that other personal or professional characteristics, beyond those considered in this study, may predict forecasting accuracy to a greater degree. For example, different ways of thinking, such as being more or less open-minded or more or less comfortable with uncertainty, may be worth exploring in future research.

Overall, our findings indicate that using an online checklist intervention did not lead to a statistically significant improvement in social workers' forecasting accuracy, as measured by Brier scores. However, the intervention group demonstrated lower scores on average, suggesting the use of a checklist may have potential in improving forecasting accuracy. Despite the lack of evidence from this study, checklists remain a practical, low-cost tool that social work teams could trial in real-world settings. In practice, teams could consider incorporating the checklist into routine supervision as a brief decision-making pause. For example, before finalising actions, social workers could identify two recommended options and one option not recommended, articulate a pro and a con for each, note at least one piece of disconfirming evidence, and complete an equity check (item 8 on the checklist). Similarly, teams could experiment with small-scale forecasting exercises by selecting two or three recurrent outcomes (such as whether a strategy meeting will occur within 12 weeks), recording forecasts monthly, and reviewing accuracy quarterly to build collective calibration. If use of the full checklist was deemed not possible, utilising just item 8, which asks whether social workers' recommendations would be the same for families of a different ethnicity or social class, serves as a simple equity check for social work teams that could be embedded into supervision as a short reflective practice. These examples are low-cost, easy to implement, and provide tangible steps that social work teams can adopt immediately to strengthen decision-making processes.

Limitations

The first limitation of this study is the delivery of the intervention. Checklist interventions are typically designed to be used within actual decision-making settings and are evaluated accordingly. Because of the COVID-19 pandemic, the checklist used in this study was designed so that it could be used online. However, similar tools would most likely be used differently in practice, for example, in discussion with a supervisor in a supervision session. Consequently, it is difficult to generalise from these findings as to what might happen if the same (or a similar) checklist were used in real-life social work decision-making environments.

Another limitation is a lack of a pre-intervention measure of confirmation bias. At the time of the study, there was no measure of confirmation bias that allowed pre- and post-testing. The Wason Selection Task was therefore used only as a post-test measure of confirmation bias, and this was unable to be compared with participants' levels of confirmation bias before using the checklist intervention. Moreover, the Wason task calls for responses to abstract logic problems, which may not adequately reflect the kind of reasoning required in social work practice. Future studies could benefit from using practice-embedded measures of confirmation bias, such as tracking responses to supervision prompts like 'What evidence would change your mind?' and quantifying how often participants actively seek disconfirming evidence.

Finally, 235 participants accessed the survey link, but only 92 met the $\geq 75\%$ completion threshold, and four were subsequently excluded as likely invalid responses. This was problematic for two reasons. The level of attrition substantially reduced the sample size and, consequently, the statistical power of the study. As a result, the study was more likely to detect moderate or large effects than smaller, yet potentially meaningful effects. Beyond reducing statistical power, the 143 incomplete responses raise concerns about the potential for self-selection bias. Participants who completed the survey may differ systematically from those who dropped out, which potentially limits the representativeness of the sample and may affect the generalisability of the findings.

Conclusions

This RCT tested an online checklist intervention to explore whether it could help improve forecasting accuracy or reduce confirmation bias for social workers, and whether confirmation bias is associated with forecasting accuracy. There was a post-intervention reduction in Brier scores for participants in the intervention group, indicating that participants' forecasting accuracy improved after the checklist intervention. However, this difference between the intervention and control groups was not statistically significant ($\beta = -0.127$, $P = .23$). The checklist intervention also

appeared to make no meaningful difference to confirmation bias ($\beta = -0.10$, $P = .38$), with similar scores obtained from the Wason Selection Task between the two groups. The personal and professional characteristics of the participants were not associated with forecasting accuracy or confirmation bias. Simply, there was no evidence that the checklist intervention made a significant difference to social workers' forecasting accuracy or confirmation bias.

A small sample size, a lack of appropriate outcome measures, and the online nature of the intervention may have influenced these results. Future research of the checklist intervention should ensure sufficient recruitment to enable more definitive conclusions to be drawn about its effectiveness. Locating the study in real-life practice settings, for example, in a supervision case discussion between a social worker and a manager, may also be more beneficial than delivering the intervention through an online vignette design.

More generally, future interventions to improve forecasting accuracy in social work should be more in-depth than the simple interventions tested in this and previous studies (Wilkins et al., 2020). In other fields, more in-depth interventions have shown marked improvements in forecasting accuracy. For example, one-hour training courses can improve forecasting ability, even among those who are already proficient at forecasting (Chang et al., 2016).

The overall aim of this and similar studies is to help establish what level of forecasting accuracy can reasonably be expected from social workers given the complexity of the practice environment, to understand more about what factors impact their ability to make accurate forecasts, and to explore the nature of the relationship between forecasting and decision-making in social work.

Acknowledgements

The CASCADE partnership receives infrastructure funding from Health and Care Research Wales.

Conflicts of interest. None declared.

Funding

This work was funded by the What Works Centre for Children's Social Care, now called Foundations—www.foundations.org.uk.

References

- Baumann, et al D. J.. (2013) 'The Decision-Making Ecology', in A. Shlonsky and R. Benbenishty (eds.) *From Evidence to Outcomes in Child Welfare: An International*

- Reader, pp. 24–40. New York: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199973729.003.0002>
- Berthet, V. (2022) 'The Impact of Cognitive Biases on Professionals' Decision-Making: A Review of Four Occupational Areas', *Frontiers in Psychology*, 12: 802439. <https://doi.org/10.3389/fpsyg.2021.802439>
- Brier, G. W. (1950) 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review*, 78: 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:vofeit>2.0.co;2](https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2)
- Catchpole, K., and Russ, S. (2015) 'The Problem with Checklists', *BMJ Quality & Safety*, 24: 545–9. <https://doi.org/10.1136/bmjqs-2015-004431>
- Chang, W. et al. (2016) 'Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments', *Judgment and Decision Making*, 11: 509–26. <https://doi.org/10.1017/S1930297500004599>
- Dalglish, L. I. (1988) 'Decision Making in Child Abuse Cases: Applications of Social Judgment Theory and Signal Detection Theory', in B. Brehmer and C. R. B. Joyce (eds) *Advances in Psychology*, Vol. 54, pp. 317–60. North-Holland: Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62178-0](https://doi.org/10.1016/S0166-4115(08)62178-0)
- Draws, T. et al. (2021) 'A Checklist to Combat Cognitive Biases in Crowdsourcing', *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9: 48–59. <https://doi.org/10.1609/hcomp.v9i1.18939>
- Drury-Hudson, J. (1999) 'Decision Making in Child Protection: The Use of Theoretical, Empirical and Procedural Knowledge by Novices and Experts and Implications for Fieldwork Placement', *British Journal of Social Work*, 29: 147–69. <https://doi.org/10.1093/oxfordjournals.bjsw.a011423>
- Elstein, A. S. (1999) 'Heuristics and Biases: Selected Errors in Clinical Reasoning', *Academic Medicine*, 74: 791–4. <https://doi.org/10.1097/00001888-199907000-00012>
- Ely, J. W., Graber, M. L., and Croskerry, P. (2011) 'Checklists to Reduce Diagnostic Errors', *Academic Medicine*, 86: 307–13. <https://doi.org/10.1097/ACM.0b013e31820824cd>
- Emmons, D. L. et al. (2018) 'Mitigating Cognitive Biases in Risk Identification: Practitioner Checklist for the Aerospace Sector', *Defense Acquisition Research Journal*, 25: 52–93. <https://doi.org/10.22594/dau.16-770.25.01>
- Featherston, R. et al. (2019) 'Interventions to Mitigate Bias in Social Work Decision-making: A Systematic Review', *Research on Social Work Practice*, 29: 741–52. <https://doi.org/10.1177/1049731518819160>
- Gawande, A. (2010) *The Checklist Manifesto: How to Get Things Right*. London: Penguin Books.
- Gigerenzer, G. and Gaissmaier, W. (2011) 'Heuristic Decision Making', *Annual Review of Psychology*, 62: 451–82. <https://doi.org/10.1146/annurev-psych-120709-145346>
- Gilovich, T., Griffin, D., and Kahneman, D. (eds) (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.
- Heick, T. (n.d.) *The Cognitive Biases List: A Visual of 180+ Heuristics*. TeachThought. <https://www.teachthought.com/critical-thinking/cognitive-biases>
- Kahneman, D. (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
- Keddell, E. (2014) 'Current Debates on Variability in Child Welfare Decision-Making: A Selected Literature Review', *Social Sciences*, 3: 916–40. <https://doi.org/10.3390/socsci3040916>

- Keddell, E., and Hyslop, I. (2019) *I'm in Two Minds about It: Decision Variability in Child Protection*. University of Otago.
- Meindl, M., and Wilkins, D. (2025) 'Can Social Workers Forecast Future Actions, Events, and Outcomes? A Study of Referrals to Children's Services in England', *Child Care in Practice*, 31: 321–35. <https://doi.org/10.1080/13575279.2021.2001434>
- Munro, E. (2011) *The Munro Review of Child Protection: Final Report – A Child-Centred System*. Department for Education.
- Murata, A., Nakamura, T., and Karwowski, W. (2015) 'Influence of Cognitive Biases in Distorting Decision Making and Leading to Critical Unfavorable Incidents', *Safety*, 1: 44–58. <https://doi.org/10.3390/safety1010044>
- Nickerson, R. S. (1998) 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises', *Review of General Psychology*, 2: 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- O'Sullivan, T. (2011) *Decision Making in Social Work*. Basingstoke: Palgrave Macmillan.
- House of Lords. (2008) *Judgements – In Re B (Children) (Fc)*, [2008] UKHL 35.
- Sanchez, C., and Dunning, D. (2018) 'Overconfidence Among Beginners: Is a Little Learning a Dangerous Thing?', *Journal of Personality and Social Psychology*, 114: 10–28. <https://doi.org/10.1037/pspa0000102>
- Saposnik, G. et al. (2016) 'Cognitive Biases Associated With Medical Decisions: A Systematic Review', *BMC Medical Informatics and Decision Making*, 16: 138. <https://doi.org/10.1186/s12911-016-0377-1>
- Tetlock, P. E., and Gardner, D. (2016) *Superforecasting: The Art and Science of Prediction*. London: Random House.
- Tversky, A., and Kahneman, D. (1974) 'Judgment Under Uncertainty: Heuristics and Biases', *Science (New York, N.Y.)*, 185: 1124–31. <https://doi.org/10.1126/science.185.4157.1124>
- Wason, P. C. (1968) 'Reasoning About a Rule', *The Quarterly Journal of Experimental Psychology*, 20: 273–81. <https://doi.org/10.1080/14640746808400161>
- Wilkins, D. et al. (2020) *Good Judgement and Social Work Decision-Making: A Randomised Controlled Trial of Brief Interventions to Improve Forecasting*. What Works for Children's Social Care. <https://orca.cardiff.ac.uk/id/eprint/132943/>
- Wilkins, D., and Meindl, M. (2021) *A Randomised Controlled Trial of a 'Checklist' Intervention to Mitigate Confirmation Bias and Improve Forecasting Accuracy in Social Work*. What Works Centre for Children's Social Care. <https://foundations.org.uk/our-work/publications/a-randomised-controlled-trial-of-a-checklist-intervention-to-mitigate-confirmation-bias-and-improve-forecasting-accuracy-in-social-work/>