GOOD JUDGEMENT AND SOCIAL WORK DECISION-MAKING: A RANDOMISED CONTROLLED TRIAL OF BRIEF INTERVENTIONS TO IMPROVE FORECASTING

January 2020

Authors

Wilkins D., CASCADE, School of Social Sciences, Cardiff University, **Foster C**., CASCADE, School of Social Sciences, Cardiff University, **Sanders M**., What Works for Children's Social Care and **Reid L**., What Works for Children's Social Care.

About the What Works for Children's Social Care

The What Works for Children's Social Care seeks better outcomes for children, young people and families by bringing the best available evidence to practitioners and other decision makers across the children's social care sector. We generate, collate and make accessible the best evidence for practitioners, policy makers and practice leaders to improve children's social care and the outcomes it generates for children and families.

About CASCADE

The Children's Social Care Research and Development Centre (CASCADE) is concerned with all aspects of community responses to social need in children and families, including family support services, children in need services, child protection, care-experienced children and adoption. It is the only centre of its kind in Wales and has strong links with policy and practice.

To find out more visit the Centre at: **whatworks-csc.org.uk**, or CASCADE at: **sites.cardiff.ac.uk/cascade**

If you'd like this publication in an alternative format such as Braille, large print or audio, please contact us at: **wwccsc@nesta.org.uk**

GOOD JUDGEMENT AND SOCIAL WORK DECISION-MAKING: A RANDOMISED CONTROLLED TRIAL OF BRIEF INTERVENTIONS TO IMPROVE FORECASTING

January 2020

EXECUTIVE SUMMARY

Social work decision making and forecasting

Child and family social workers make many decisions every day, a significant number of which have the potential to be life changing for the children and families involved. A high proportion involve explicit or implicit forecasts about what may happen in the future, for example what will happen if we decide to accept the referral or close the case and what sort of 'intervention' will lead to what sort of outcome (Cartwright and Hardie, 2017). In this study, we explore two key questions – first, how accurately can social workers forecast what is going to happen next from reading brief case studies similar to social work referrals? Second, can the accuracy of these forecasts be improved via brief, online interventions? In future studies, we aim to explore the relationship (if any) between more accurate forecasting and outcomes in the context of children's services. The study is inspired by and aims to build upon the methods and success of the **Good Judgement Project** (GJP; Tetlock and Garder, 2016).

Study design and sample

283 participants took part in an online survey in which they were asked to read four case studies based on real referrals to children's services. After each case study they were asked to estimate the likelihood of different outcomes. After the first two case studies, participants were randomly assigned to take part in one of three very brief online interventions or a control condition (see figure 1). The three interventions were confidence calibration, cognitive debiasing and growth mindset with feedback. The first involved answering ten questions about social work and indicating a level of confidence for each answer. Participants then received feedback on whether they were under-confident, over-confident or well-calibrated. The second involved reading background information about confirmation bias and being asked to consider how another individual might consider the same referral information and arrive at a different conclusion. For the third intervention, participants were given the answers for the first two case studies, so that they could compare with their own responses.

Results

1. How accurately can social workers predict outcomes?

On average the survey participants were only slightly better than chance at anticipating the outcome for each referral. Some of the case studies were more difficult to forecast than others. The only individual

characteristic related to forecasting accuracy was age (and not gender, ethnicity, current role, current team or length of post-qualifying experience) – with younger participants being on average more accurate than older participants.

2. Can brief online interventions improve the accuracy of predications?

None of the interventions resulted in significantly better forecasts, suggesting that - at least in this study - these brief interventions were ineffective.

Implications

These findings suggest that, for most social workers, it is very difficult to predict what might happen in a given case based on scant information. However, a small number of participants in the study were able to make very accurate forecasts. This suggests that it is possible to make accurate predictions based on scant information, although further study would be needed to establish the extent to which high (and low) performing participants revert to the mean over time. Further study is also needed to understand the relationship between more accurate forecasting and outcomes for children and families, as well as to develop more in-depth methods to improve forecasting accuracy.
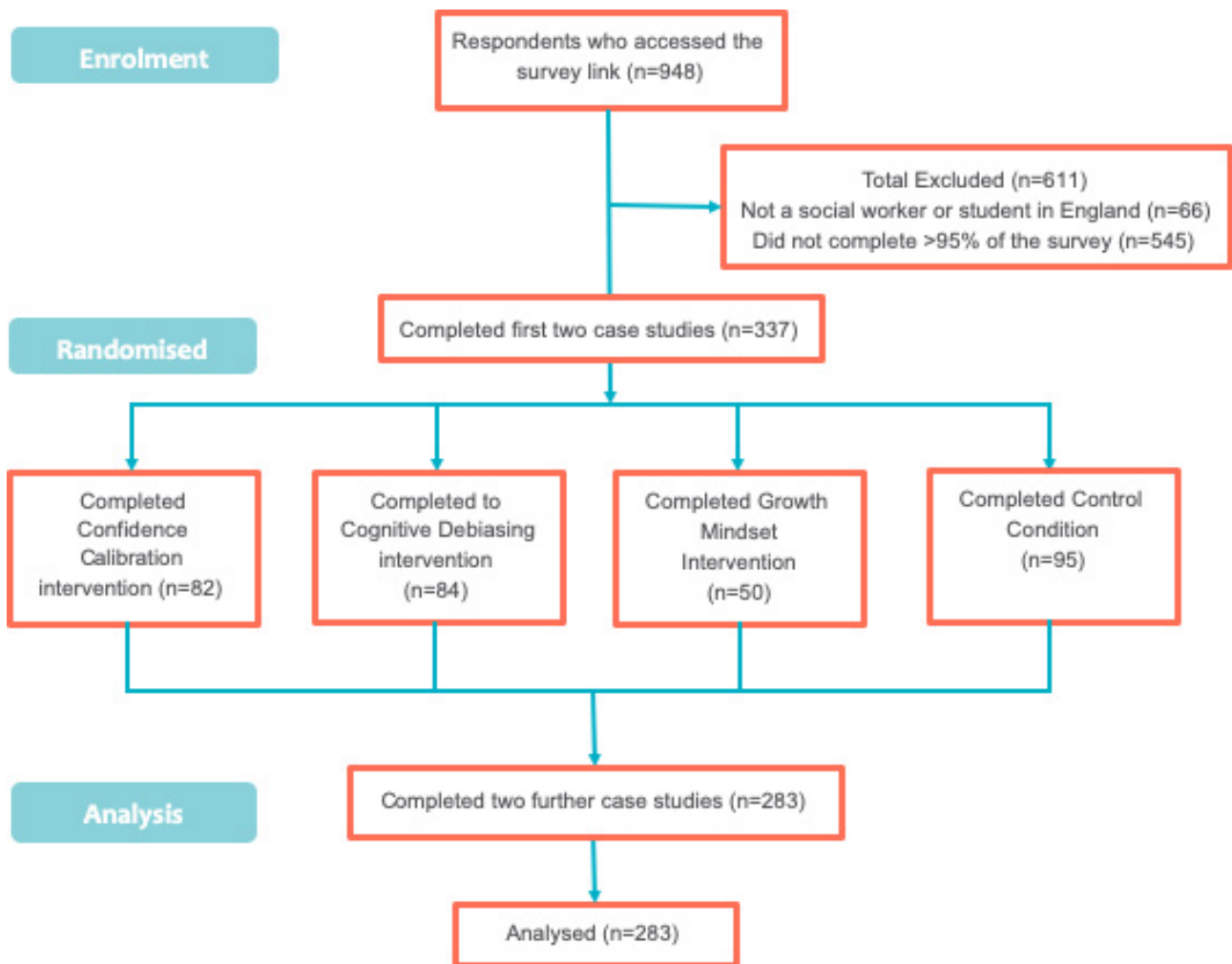
Figure 1. Pathway of the RCT from enrollment to analysis

INTRODUCTION

This report describes a randomised controlled trial (RCT) of three very brief online interventions to improve decision-making skills. The RCT was undertaken in the form of an online survey aimed at student and qualified social workers in England and ran between August and October 2019. The three interventions tested were confidence calibration, cognitive debiasing and growth mindset with feedback. The first involved answering ten questions about social work and indicating a level of confidence for each answer. Participants then received feedback on whether they were under-confident, over-confident or well-calibrated. The second involved reading background information about confirmation bias and being asked to consider how another individual might consider the same referral information and arrive at a different conclusion. For the third intervention, participants were given the answers for the first two case studies, so that they could compare with their own responses.

In the survey, participants were asked to read four case studies, all of which were based upon actual referrals made to children's services between late 2018 and early 2019. After each case study, participants were asked four questions about the likelihood of different outcomes and their estimates compared with the actual outcomes of the referrals. Between referrals two and three, participants were randomly assigned to one of three brief interventions or a control condition. After the fourth referral, participants were asked a series of questions about their personal and professional characteristics and about their views of child and family social work.

Background and rationale

Child and family social workers in England make thousands of decisions every day, many of which have the potential to be life changing. These decisions are made in the context of the wider system (Munro, 2005) and are inevitably influenced by macro-level inequalities, as well as organisational and case-related factors (Keddell and Hyslop, 2016). Ideally, the majority of these decisions will be made in partnership with families, including parents and young people. This complexity can create a challenging decision-making environment for which social workers need to have (or develop) fundamental decision-making skills such as open-mindedness, cognitive reflection, pattern recognition and creative problem solving, alongside their relationship-based skills of practice (Featherston et al., 2019).

An emerging body of work on making good judgements in conditions of uncertainty offers some potentially new and exciting insights that could be used to help social workers examine and improve their decision-making abilities. Specifically, the Good Judgement Project (GJP) have demonstrated how the 'wisdom of crowds' (Surowiecki, 2004) can be utilised to improve upon the judgement of individuals. Using a series of interventions, including screening for existing decision-making skills, online and in-person training sessions, structured group work and feedback, the GJP has been able to significantly improve the accuracy of future forecasts (Tetlock et al., 2014, Ungar et al., 2012). This is notable, given that most people over-estimate their forecasting abilities and rarely perform better than chance in practice, this being especially true over extended time periods. In this study, we explored whether social workers and student social workers could answer forecasting questions in relation to case studies based on real-life referrals and whether three brief online interventions would improve the accuracy of these forecasts.

METHODS

Research questions

Our primary research questions were:

(1) How accurately can social workers forecast what is going to happen next from reading case studies based on real-life referrals? and,

(2) Can the accuracy of social work forecasting be improved using brief online interventions?

We also sought to address the following secondary research question:

(3) Can we predict forecasting accuracy from personal or professional characteristics?

Ethics

The School of Social Sciences' Research Ethics Committee at Cardiff University gave approval for the study. At the outset of the survey, participants were provided with an information sheet and asked to sign a consent form, as part of which they were informed that taking part was voluntary and they could exit the survey at any point. All data were collected anonymously, unless participants opted to provide their email address (to signal that they would be interested in taking part in further such studies).

Sampling

Social workers and student social workers in England were recruited via the What Works for Children's Social Care (WWCSC) website and social media feeds, through direct approaches to partner local authorities and other social work organisations including Frontline. Before completing the survey, participants were asked to declare whether they were either a qualified social worker or student social worker in England. A negative response to this question did not prevent the participant from taking part but did allow us to separate the sample into qualified and student social workers and other participants. All results presented in this report relate only to participants who self-declared as a social worker or student social worker in England. A £5 charitable donation was promised by the WWCSC on behalf of each participant who completed the survey, with an extra donation of £2 to be made if the participant's overall accuracy was better than chance. These donations have been shared equally between the charities **Trevi House** and **Become**. The performance-based element of the donation was used to incentivise participants to take due care over their responses.

The case studies were selected by the first author from a list provided by one local authority in England of referrals received between January and March 2019. The list was sub-divided into four categories, based on the outcome of the referral: (1) no further action taken, (2) closed following a social work assessment, (3) child made subject to a child in need or child protection plan, and, (4) the child came into care. The lead author selected at random three referrals from each list and then read the case file to establish what questions might be asked about each one (in addition to the primary outcome question).

Data collection

The survey was hosted in Qualtrics (**https://www.qualtrics.com**) and available for participation between 16th August and 31st October 2019. As outlined above, the survey consisted primarily of four case studies (the first two of which were the same for all the participants and the second two drawn from a larger pool of ten cases at random for each participant). In relation to each one, participants were asked four questions about the likelihood of different next steps and outcomes, for example:

- In response to this referral, how likely is the local authority to take 'no further action'?

- Within the next 12 weeks, how likely is it that the child will come into care?

- Within the next 12 weeks, how likely is it that the mother will attend at least one session of a parenting programme?

- Within the next 6 months, how likely is it that another referral will be made about the same child?

Participants were asked to respond using a scale from 0 - 100, where 0 means definitely will not happen and 100 means definitely will happen.

As the case studies were based on real referrals, we had already obtained the answers to these questions and could therefore compare participant responses with what actually happened. In between case studies two and three, participants were randomised to one of three brief interventions or a control condition. At the end of the survey, participants were asked questions about their professional and personal characteristics (e.g. age, gender, current role, length of experience) and about their views of children's services (e.g. what is the main purpose of children's services?). All participants completed the same two case studies at

baseline. The post-intervention third and fourth case studies were selected at random for each participant from a group of ten.

Description of the three interventions

Confidence Calibration

Most people are either overconfident, they think they know more than they do, or underconfident, they think they know less than they do (Hattie, 2013). Particularly for novices and near experts, it can be hard to judge accurately your own level of knowledge. In this intervention, participants were asked to respond with true or false to ten social work-related factual statements. For example: *There was a 4% increase in children in need in England between 31st March 2017 and 31st March 2018.*

As well as a true or false response, participants were asked to indicate how confident they felt about their answer (from 50 to 100 per cent sure). After completing the ten statements, participants were shown how many they answered correctly and their average level of confidence. The aim of the intervention was to show participants whether they were well calibrated, overconfident or under confident, in order that they might adjust their answers for the next set of case studies accordingly.

Cognitive Debiasing

Human reasoning is subject to a number of systematic biases (Tversky and Kahneman, 1974). For this intervention, participants were provided with background information about confirmation bias, before being asked to consider how their beliefs about children's services might influence their analysis of the case studies (De Haal et al., 2019; Morewedge et al., 2015). The aim of the intervention was to help participants consider the case studies from more than one point of view by thinking about how someone with a different set of beliefs might see things differently.

Growth Mind-Set and Feedback

Providing people with feedback on their performance is a widely used method for improving performance in a range of different situations (Balcazar et al., 1985). How receptive people are to feedback may be related to their mindset in relation to learning – whether they have a 'growth mindset' (open to feedback) or a 'static mindset' (closed to feedback). In this intervention, participants were given background information about the importance of feedback and then provided with the actual outcomes of the baseline case studies. The aim of the intervention was to show participants where they might have been wrong (or right) in

relation to the first set of questions, so that they could adjust their answers for the next two case studies accordingly.

Control

In the control condition, participants were asked to read some generic information about the importance of reflective practice.

Data analysis

Data analysis was carried out in **RStudio** and **JASP.** All R code can be found at **https://github.com/ CatherineFoster**. Participants were excluded from the analysis if they did not declare themselves to be a registered social worker or student social worker in England, had not completed the full survey (>98%) or had a response time of <5 minutes. The first step was to calculate Brier scores for each individual. Brier scores are used to measure the accuracy of probabilistic predictions and range from 0 to 2, where 0 indicates complete accuracy and 2 indicates complete inaccuracy. As a benchmark, a Brier score of 0.5 can be considered equal to chance (the score you would obtain if forecasting the outcome of a series of coin tosses). Brier scores are calculated using the following formula, where x = the forecast for the outcome that occurs and y = the forecast for the outcome that does not occur.

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

Mean baseline Brier scores were calculated using the eight questions relating to the first two case studies. Mean post-treatment Brier scores were calculated using the eight questions for the third and fourth case studies. An overall mean Brier score was calculated using the answers to all sixteen questions to determine overall accuracy and conduct exploratory analysis. An important feature of Brier scores is that lower scores indicate greater accuracy.

Post-intervention Brier scores were used to determine whether the interventions were effective in relation to improved forecasting accuracy. Baseline Brier scores and post-intervention case studies completed were included in the regression model as covariates, the latter to control for differences between the cases. The correction for differences in accuracy at baseline was performed because it was assumed that the groups to be compared were equal at baseline. The post-intervention situation is different, because differences between the groups could be caused by the fact that each group received a different intervention (or control condition).

RESULTS

283 student and qualified social workers in England completed ≥98% of the study and were included in the full analysis. Outliers were not removed. Not all participants provided answers for every question.

Participant demographics

Of the 282 participants who stated their ethnicity, 87% were White (from the UK, Ireland or Other), 2% were White and Asian, 2% White and Black Caribbean, 1% Mixed Ethnic Background, 1% Asian British Indian, 1% Black British Caribbean, 1% Black British African and 1% Other Ethnic Group. The remaining ethnicities each made up < 1% of the sample. The largest age group was 25 to 34 years old, while more than four-fifths of the sample were female (figure 2). The mean level of post-qualifying experience within the sample was varied, with a mean of 8.88 years (sd=8.6). The largest groups in terms of current role were student social worker or social worker (figure 3).

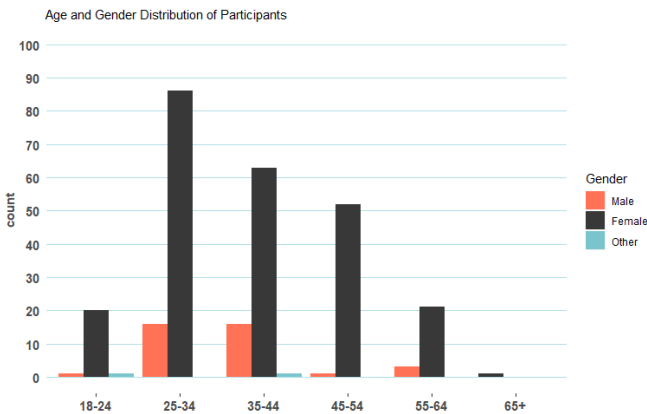### THE MAJORITY OF PARTICIPANTS WERE FEMALE AGED 25-34

Figure 2. N=283. 88% of the sample was female and 36% aged 25-34. 0.7% of the sample (n=2), listed their gender as Other.

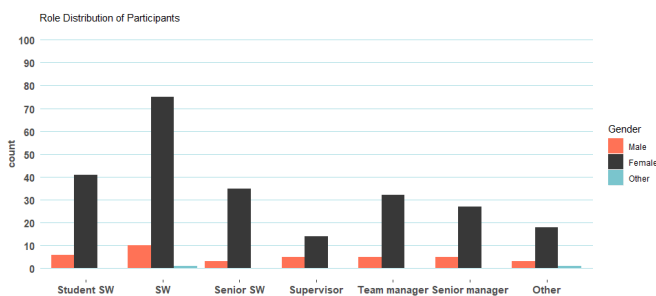### THE HIGHEST PROPORTION OF PARTICIPANTS WERE QUALIFIED FEMALE SOCIAL WORKERS

Figure 3. N=282. 30% of the sample were social workers and 17% student social workers. 8% chose "other" but either confirmed that they are currently registered as a social worker in England or provided a job title within social work. SW = Social worker.

Most of the sample either had obtained or were studying for a BA/BSc qualification, albeit this group was only modestly larger than the MA/MSc group. The majority of participants were currently studying or had qualified via traditional rather than fast-track courses (figure 4).

90% OF THE SAMPLE HAD AT LEAST ONE DEGREE AND QUALIFIED THROUGH TRADITIONAL ROLES
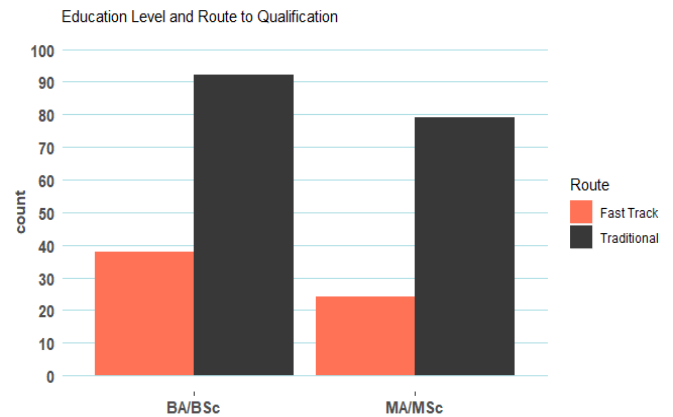

Figure 4. N=233 provided route to qualification information. Of the n=282 who provided education level data, 50% were educated to university undergraduate level, and 40% to MA/MSc level. Although we are not aware of any BA/BSc fast-track courses in England, this could refer to e.g. completing the Frontline programme and obtaining a Postgraduate diploma in the first year, without completing the second year and obtaining an MA.

Overall accuracy

The mean Brier score for the full sample in relation to all baseline and post-treatment questions was 0.47 (sd=0.13), and 177 of the 283 participants scored above chance, i.e. 0.49 or better (figure 5). The remaining participants (n=106) were equal or worse than would be expected by chance. A one-way ANOVA confirmed that there were no baseline differences in Brier scores between the four conditions; $F(3,279)=2.363$, $p=0.072$ (Table 1).

Summary of Baseline and Post-Treatment Brier Scores										
	Control N=94		Cognitive Debiasing N=69		Confidence Calibration N=74		Growth Mindset N=47		Total group N=283	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Brier Score (Baseline)	0.44 (0.17)	0.43 (0.22)	0.41 (0.17)	0.38 (0.21)	0.38 (0.14)	0.37 (0.17)	0.39 (0.12)	0.4 (0.17)	0.41 (0.16)	0.4 (0.2)
Brier Score (Post-Intervention)	0.54 (0.16)	0.53 (0.17)	0.51 (0.13)	0.50 (0.17)	0.55 (0.18)	0.54 (0.2)	0.52 (0.15)	0.52 (0.19)	0.54 (0.16)	0.52 (0.18)
Pre-Post Difference in Brier Score	0.1 (-0.01)	0.1 (-0.01)	0.1 (-0.04)	0.12 (-0.4)	0.17 (0.04)	0.17 (0.3)	0.13 (0.03)	0.12 (0.02)	0.13 (0)	0.12 (-0.02)

Participants were more accurate in relation to the first two baseline case studies than they were for the post-intervention case studies. 74% of participants were more accurate than chance for the former, and only 42% for the latter.

As the post-treatment case studies were randomised, with each participant seeing two out of ten possible cases, we examined the mean Brier scores for each one (figure 4). The cases with the least accurate Brier scores were Case 1 (mean=0.73, sd=0.28) and Case 4 (mean=0.68, sd=0.1). Case 7 had the most accurate Brier scores (mean=0.34, sd=0.15).

THERE WAS VARIABILITY IN FORECASTING DIFFICULTY BETWEEN THE CASES
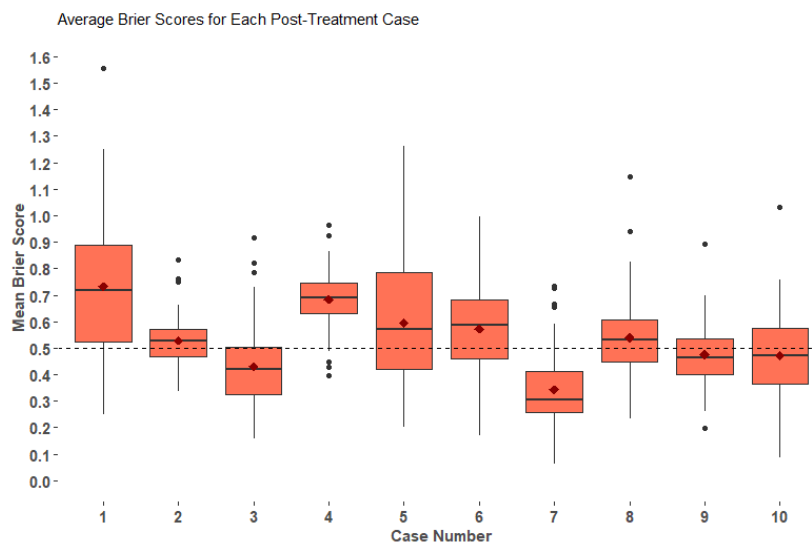

Average Brier Scores for Each Post-Treatment Case

Figure 5. N=283. Red dot indicates mean Brier score and central black band indicates the median. Black dots represent outliers. Dashed line represents Brier scores at chance-level.

Intervention effects

Baseline and post treatment Brier scores were moderately correlated ($r(281)=.35$, $p<0.001$; figure 6). Baseline scores were included in the regression analysis of intervention effects.
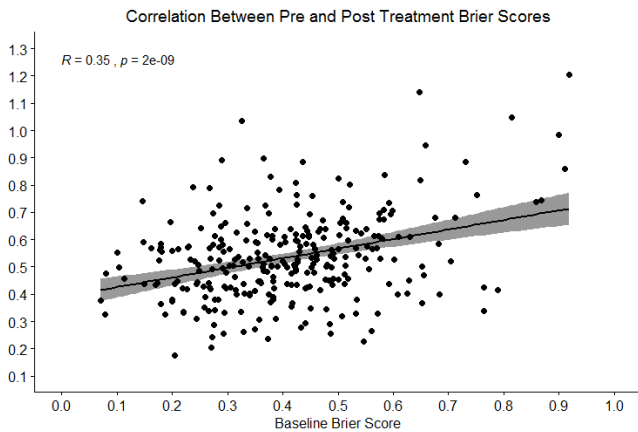
Figure 6. N=283. Scatterplot showing the correlation between baseline and post-treatment Brier scores for the full samples. Shaded area represents the 95% confidence interval on the correlation.

Via a regression analysis we compared each intervention to the control condition, with baseline Brier scores and case studies viewed post-intervention included as covariates. Non-influential variables of age, gender and ethnicity were removed from the final regression model, as they each explained less than 2% of the variance. Given the low frequency of ethnicities other than White (87% of total sample), ethnicity was also removed from the exploratory analysis.

The full model explained 24.4% of the variance in the data, $r2 = 0.244$, $F(14,268)=7.5$, $p<.001$). No intervention performed better than the control condition. The confidence calibration intervention resulted in the least accurate Brier scores of any group, with an increase of 0.03 on average compared to the control condition (figure 7).
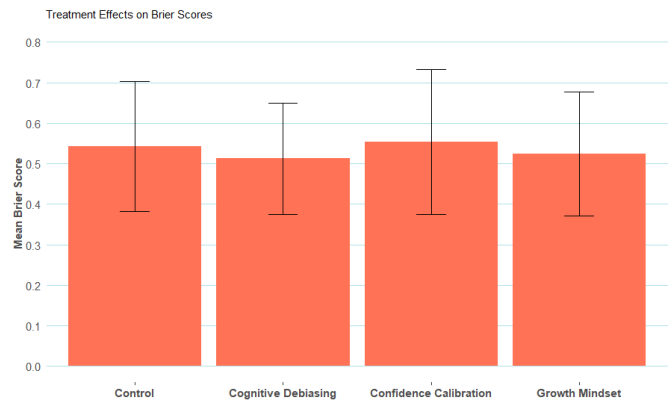
Figure 7. N=283. Barplot showing the mean post-treatment Brier scores for each group, error bars represent standard deviation from the mean. The cognitive de-biasing intervention group had more accurate (lower) Brier scores than all other conditions, however this difference was not significant in the regression, $p > 0.05$.

The post-intervention cases seen by participants affected their post-treatment Brier scores. Participants who saw Case 1 or Case 5 had significantly higher (less accurate) Brier scores ($p<0.001$, and $p=0.006$ respectively), and those who saw Case 7 had significantly lower (more accurate) scores ($p=0.002$). Case 3 also resulted in marginally better scores ($p=0.04$). See supplementary Table 1 for the full results from the regression.

Exploratory analysis

Additional exploratory analyses were conducted in order to generate hypotheses for further study. For this, we used personal and professional characteristics data and views as to the purpose of children's services.

First, we developed a second regression model to examine demographic and current role effects on Brier scores. The variables included in this regression were: age, gender, current role, current team, length of post-qualifying experience, route to qualification and highest level of education. Ethnicity was removed as the majority of the sample identified as White. Many of these variables were correlated with Brier scores. The variance inflation factor (VIF) which identifies correlation between independent variables and the strength of that correlation was examined to check for issues of multicollinearity. VIFs<5 are not considered strong enough to warrant corrective steps, the highest VIF in the exploratory correlation was between age and years of experience, VIF=2.6, therefore, no variables were removed.

The exploratory model explained 3.1% of the variance in the data (adjusted $r2=.003$, $F(7,243)=1.11$, $p=0.357$). No variable significantly predicted Brier scores, although age was further examined as there was some indication of a relationship with Brier scores ($t=1.7$, $p=0.08$)

Age alone was a significant predictor of Brier scores when included as the only variable in the model ($r2=.0020$, $F(1,282)=5.749$, $p=0.017$). Those in the age group 25 to 34 years achieved the lowest (most accurate) Brier scores (mean=0.44, sd=0.09), followed by those in the age group 18 to 24 years (mean=0.5, sd=0.16). Following these groups, there was an inverse relationship between age and forecasting accuracy (figure 8).
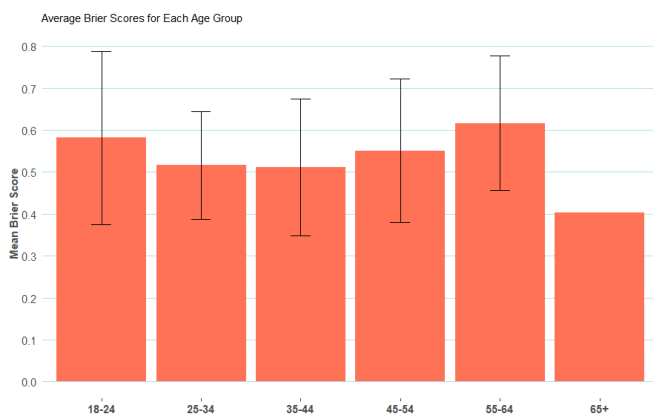
Average Brier Scores for Each Age Group

Figure 8. N=283. Brier scores increased with age after the age bracket 25-34. Mean scores shown, error bars represent standard deviation from the mean.

Beliefs and forecasting accuracy

We then examined the effect of responses to the following statements, scored using a five-point Likert scale from 1 (Strongly agree) to 5 (Strongly disagree):

(1) The main purpose of children's services is to support parents

(2) The main purpose of children's services is to protect children from abuse and neglect

(3) The main purpose of children's services is to protect human rights within a liberal society

Participants were asked to respond to all three statements, and they were not mutually exclusive (e.g. participants could strongly agree or disagree with all of them). The only belief significantly associated with forecasting accuracy was statement 1 (figure 8).

Those who strongly disagreed provided significantly less accurate forecasts than those who chose any other response (figure 9). ANOVA showed that Levene's test of equality of variances was violated, $F(4, 278)=2.93$, $p=0.021$, therefore the Brown-Forsythe correction was applied as the data had a non-Gaussian distribution, $F(4, 21)=3.32$, $p=0.021$. Post-hoc tests showed that participants who chose scale point 5 were less accurate than those who chose any other scale point (all $p\geq0.001$). However, as only seven participants chose a point 5 response, compared with 25 for point 4, 42 for point 3, 146 for point 2 and 63 for point 1, the validity of this finding would need to be tested in a larger sample.

PARTICIPANTS WHO STRONGLY DISAGREE THAT THE MAIN ROLE OF CHILDREN'S SERVICES IS TO SUPPORT PARENTS MADE THE LEAST ACCURATE FORECASTS

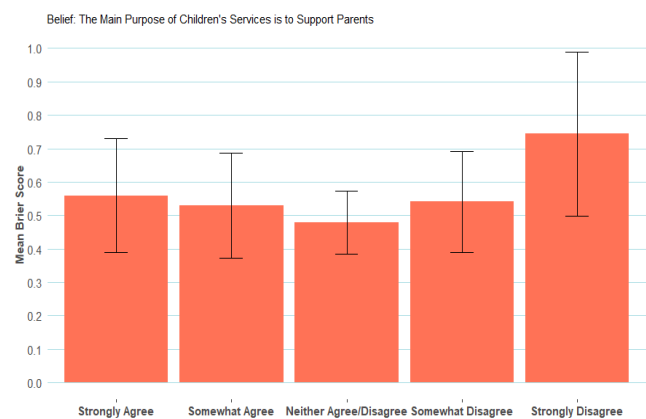Belief: The Main Purpose of Children's Services is to Support Parents

Figure 9. N=283. Participants who strongly disagreed (N=7) that the core role of children's services is to support parents were significantly less accurate than those who strongly agreed with this statement. Mean Brier score shown for each group, error bars represent standard deviation. Scale: from 1= Strongly agree to 5=Strongly disagree.

Strengths and limitations

Strengths of the study include the relative diversity of the sample, in relation to geography, team and current roles and levels of experience. The anonymous design also means there is no risk of 'groupthink.' As well-trained and structured groups tend to provide more accurate forecasts than individuals, strategies for avoiding the negative consequences of groupthink are an important part of forecasting training (Mellers et al, 2014), albeit this is outside the direct scope of this study.

The study also has several limitations. A methodological limitation is that we cannot compare the difficulty

of cases before and after the interventions. While we can compare the Brier scores for each case post-intervention, it is not clear whether the more accurate Brier scores for the baseline case studies were because they are easier to forecast, because the interventions negatively affected performance or because performance simply degraded over time for other reasons (e.g. fatigue). We should have randomised the presentation of all the case studies, not only those post-intervention and would do so in any future study of similar design. The study was also underpowered for the interventions, and a larger sample would be needed in order to draw firmer conclusions about the relative status of the three interventions compared to the control condition. The case studies were also drawn from one particular local authority, and so it is possible and even likely that in other local authorities, different outcomes could have occurred. This limitation is not as severe as it might first seem, as the aim of the task was to forecast what did happen – more accurate forecasts would be achieved by those participants who took into account the unknown information (such as which local authority received the referral) as well as known information (such as that contained within the case study).

CONCLUSION

The findings from this study show how difficult it is for social workers to say with confidence what is going to happen next in relation to a selection of 'real world' referrals. The three very brief interventions tested in this study did not make a positive difference. Nevertheless, now that we have a baseline understanding of social work forecasting, there is potential to learn from the wider forecasting literature about other interventions that may prove more useful.

One noteworthy finding is the variability in Brier scores between the different case studies. This could suggest that at least some of the variability in forecasting accuracy is related to the nature of the issues involved (e.g. whether the concerns relate to domestic abuse or parenting or substance misuse, etc.) and/or other features of the case studies. It would be helpful to explore in future studies what types of issues or scenarios workers find more difficult to forecast. If we can identify the issues that are more difficult we can target efforts to improve decision-making more specifically in those areas.

Further studies are currently being planned. First, we plan to develop a more in-depth intervention in relation to cognitive debiasing aimed at a small group of social workers who we can work with more intensively.

Second, we are designing an RCT-in-the-field where in-depth forecasting training will be given to groups of social workers and forecasts made about their current caseloads, compared to forecasts made by other groups of workers without the training.

Finally, and in the longer-term, we aim to explore the relationship between improved forecasting skills and real-world decision-making and outcomes for children and families.

SUPPLEMENTARY DATA

Supplementary Table 1. Regression coefficients for each condition vs. control. Negative t-values indicate lower Briers scores (more accurate forecasts) than the control condition.

Regression Coefficients			
	Estimate	Std. Error	t-value	p-value
Intercept	0.41	0.44	9.15	< 0.001
Cognitive Debiasing	-0.02	0.02	-0.97	0.33
Confidence Calibration	0.02	0.02	0.74	0.46
Growth Mindset	-0.01	0.03	-0.41	0.68
Baseline Brier Score	0.33	0.05	6.03	< 0.001
Case 1	0.1	0.02	4.3	< 0.001
Case 2	0.002	0.23	0.09	0.92
Case 3	-0.04	0.02	-1.84	0.06
Case 4	0.06	0.02	2.7	0.007
Case 5	0.003	0.01	0.26	0.79
Case 6	0.04	0.02	1.53	0.12
Case 7	-0.03	0.01	-2.8	0.005
Case 8	0.01	0.008	1.32	0.19
Case 9	-0.03	0.02	-1.26	0.21
Case 10	-0.005	0.01	-0.5	0.61

REFERENCES

Balcazar, F., Hopkins, B. and Suarez, Y. (1985). A Critical, Objective Review of Performance Feedback. Journal of Organizational Behavior Management, 7(3-4), 65-89.

Cartwright, N. and Hardie, J. (2017). Predicting What Will Happen When You Intervene. Clinical Social Work Journal, 45, 270-279.

De Haal, W., van Berkel, S., van der Asdonk, S., Finkenauer, C., Forder, C., van IJzendoorn, M., Schuengel, C. and Alink, L. (2019). Out-of-home placement decisions: How individual characteristics of professionals are reflected in deciding about child protection cases. Developmental Child Welfare, 1(4), 312-326.

Hattie, J. (2013). Calibration and Confidence: Where to next? Learning and Instruction, 24(1), 62-66.

Featherston, R., Shlonsky, A., Lewis, C. Luong, M., Downie, L., Vogel, A., Granger, C., Hamilton, B. and Galvin, K. (2019) Interventions to Mitigate Bias in Social Work Decision-Making: A Systematic Review. Research on Social Work Practice, 29(7), 741-752.

Keddell, E. and Hyslop, I. (2016). First findings from phase one of the Child Welfare Decision-Making Variability Project. Research Briefing Paper. Otago, University of Otago.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S., Moore, D., Atanasov, P., Swift, S., Murray, T., Stone, E. and Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. Psychological science, 25(5), 1106-1115.

Morewedge, C., Yoon, H., Scopelliti, I., Symborski, C., Korris, J. and Kassam, K. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. Evaluating and Mitigating Risk, 2(1), 129-140.

Munro, E. (2005). Improving practice: Child protection as a systems problem. Children and Youth Services Review, 27(4), 375-391.

Tetlock, P. and Gardner, D. (2016). Superforecasting: The Art and Science of Prediction. London, Random House Books.

Tetlock, P., Mellers, B., Rohrbaugh, N., and Chen, E. (2014). Forecasting tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. Current Directions in Psychological Science, 23(4), 290-295.

Tversky, A. and Kahneman, D. (1974). Judgement under Uncertainty: Heuristics and Biases. Science, 185(4157), 1124-1131.

Surowiecki, J. (2004). The Wisdom of Crowds. Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Societies and Nations. New York, Random House.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., and Baron, J. (2012). The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions. 2012 AAAI Fall Symposium Series.