Journal of Social Work Practice

Psychotherapeutic Approaches in Health, Welfare and the Community

ISSN: 0265-0533 (Print) 1465-3885 (Online) Journal homepage: www.tandfonline.com/journals/cjsw20

Unwanted variability in child welfare decision-making: an empirical investigation

Lene Caroline Ljosland Nordmo, David Wilkins & Magnus Nordmo

To cite this article: Lene Caroline Ljosland Nordmo, David Wilkins & Magnus Nordmo (27 Nov 2025): Unwanted variability in child welfare decision-making: an empirical investigation, Journal of Social Work Practice, DOI: 10.1080/02650533.2025.2589778

To link to this article: https://doi.org/10.1080/02650533.2025.2589778

Published online: 27 Nov 2025.

JOURNAL OF SOCIAL WORK PRACTICE
https://doi.org/10.1080/02650533.2025.2589778

OPEN ACCESS Check for updates

Unwanted variability in child welfare decision-making: an empirical investigation

Lene Caroline Ljosland Nordmo[a], David Wilkins[b] and Magnus Nordmo[c]

[a]Inter-Municipal Resource Team Child Welfare Services Telemark County, Norway; [b]School of Social Sciences, Cardiff University, Cardiff, UK; [c]Department of Educational Science, University of South-Eastern Norway, Notodden, Norway

ABSTRACT

Assessing a parent's ability to provide safe, nurturing care is critical yet difficult. Child welfare professionals must make such judgements, but they are complex and prone to noise, understood as unwanted variability. Noise arises when decisions rely on ambiguous information and discretion-laden criteria, a common feature in child welfare. In Study 1 ($N = 49$), professionals independently rated the overall quality of care in ten vignettes varying in concern severity, information completeness, and source. Inter-rater reliability was .50, with 24.7% of the variance due to systematic rater differences, suggesting varying thresholds for unsafe parenting. In Study 2 ($N = 44$), we tested whether averaging domain-specific evaluations could reduce noise. Although reliability improved modestly, the change was not significant. These results highlight the need for noise-reduction strategies to improve consistency in child welfare assessments. Because noise is measurable and predictably introduces error, we recommend integrating targeted noise-reduction measures.

ARTICLE HISTORY
Received 22 April 2025
Accepted 31 October 2025

KEYWORDS
Child welfare; inter-rater reliability; professional judgement; signal to noise ratio; risk assement; child of impaired parents

Child protection agencies play an essential role in safeguarding children's rights and ensuring their well-being. This dual mandate involves protecting children from serious harm while also being cautious not to intrude excessively into private family life. This delicate act of balancing requires utmost professionalism in every aspect of judgement, understood here as the reasoned weighing of information, values, and uncertainty to reach a defensible conclusion (Taylor, 2012; Taylor & Whittaker, 2018). A significant challenge for child welfare service agencies is discerning between children genuinely in need of protection and those who are not. Two primary errors can occur in this process – intervening when unnecessary or overlooking a genuine need for intervention (also known as false positives and false negatives). In the field of child welfare services, the principle of erring on the side of safety, commonly applied in other risk management contexts, is complicated by the fact that both intervention and non-intervention carry distinct risks and potential harms. While the risks of non-intervention may disproportionately affect the child

CONTACT Magnus Nordmo magnus.nordmo@usn.no

and the reputation of the profession, the harms of unnecessary interventions fall heavily on families including children and particularly those from marginalised groups. This dynamic helps sustain a bias towards avoiding errors of non-intervention over those of intervention (Wilkins & Meindl, 2024). The professional judgement landscape in child protection is also notably intricate. Numerous factors come into play, each accompanied by its own set of uncertainties. Every decision must be definitive, consolidating a broad range of elements into a singular action. In such complex scenarios, variability in judgements is bound to emerge, as child welfare workers are compelled to rely to a greater or lesser extent on their intuitions as well as procedures and evidence (Munro, 1999).

Several investigations have shown that ensuring consistency and accuracy in child welfare decision making represents a formidable challenge (Benbenishty et al., 2015; Keddell, 2017b). In their review, Hood et al. (2022) summarise research on child welfare decision making into five categories: Accuracy (whether a judgement is correct given some corroborating evidence), consistency (whether judgements are consistent across raters and time), outcomes (whether judgements lead to improved outcomes), practice (whether a judgement is in agreement with best-practice procedures) and equity (whether judgements are unfairly impacted by information regarding social or demographic group membership). Child welfare decision making can also be understood in terms of statistical theory where errors are either due to bias or noise (Lwin & Beltrano, 2022). A biased judgement is when many evaluators converge on an erroneous judgement. A forecasting study allows researchers to assess whether a given forecast matches with reality after the fact. For example, you can ask social workers to predict whether a child will go into care, or whether a closed referral will be re-opened within a defined period and compare their predictions with the outcome to calculate accuracy and measure bias (Meindl & Wilkins, 2021). Noise, on the other hand, refers to the spread of the distribution of judgements. Judgements are noisy if there is variability in ratings that should be identical. For example, Rustad et al. (2022) found substantial variability in the threshold for investigating referrals to child welfare services between different Norwegian counties. A vignette study from Spain (Mosteiro et al., 2018) found that although child protection professionals drew on similar categories of arguments when justifying their decisions, they differed substantially in how much weight they attributed to each criterion and how they interpreted them. As a result, the same case description led 62% to recommend family preservation and 38% to recommend family separation, illustrating how variability emerges not only from case factors but also from professionals' thresholds and underlying beliefs about intervention outcomes. Similarly, a mixed-methods study from New Zealand (Keddell, 2017a) compared 'risk-averse' and 'risk-friendly' practitioners and found that, despite drawing on the same knowledge base and identifying similar risk and safety factors, their perceptions of harm differed markedly. Risk-averse practitioners were more certain in their conclusions, rated children's safety lower, and anticipated more severe future harm if no intervention occurred, whereas risk-friendly practitioners adopted a more tentative stance, focusing on immediate stressors rather than long-term trajectories. In a series of publications, Wilkins and Meindl (2021, 2022, 2023, 2024) explored the forecasting accuracy of UK child welfare workers by examining their predictions about the likelihood of various future actions and events. The studies overall conclusion is that average forecasts are close to chance-level, the theoretically lowest accuracy for random forecasts. In addition, child welfare workers tend to overestimate the likelihood of adverse outcomes, perhaps to prioritise the avoidance of missing genuine cases of abuse, thus erring on the side of false positive errors over false negatives (Wilkins & Meindl, 2023, 2024). This research is based on vignette studies, which involve creating brief, anonymised descriptions of hypothetical cases for analysis.

One double-edged feature of forecasting studies is the use of real-life outcomes as a measuring stick. Given the ubiquity of noise (Kahneman et al., 2021), while there is inherent validity in assessing forecasts made by child welfare workers in relation to a court decision or a clinical assessment, the measuring stick itself is unreliable. A crime might not be reported, a judge might not evaluate an action as a crime, and domestic abuse might be concealed (Austin & Thomas, 1977). This is well-documented in many aspects of the legal system, such as UK assessment of asylum seekers (Burridge & Gill, 2017). Researchers and practitioners face a fundamental challenge in evaluating bias in child welfare decisions, as absolute truth is often unknowable and, in some cases, may not exist objectively. In contrast, noise is easier to measure, objectively true, and always a source of error. A recent study shows that variability between different child welfare workers is a key component of erroneous decisions (Lwin et al., 2024). Emphasising noise, rather than forecasting alone, allows researchers and practitioners to improve processes that extend beyond the inherent uncertainty of prediction to the equally important challenges of interpreting information correctly, structuring assessments, and planning systematic information gathering.

In this paper, we quantify the amount and nature of statistical noise by comparing the assessment of several professional child-welfare workers with identical information. We aim to answer the following research questions. First, how much noise is present in the assessments of realistic, yet fictitious vignettes made by Norwegian child welfare workers? Vignette studies involve presenting respondents with standardised case descriptions that simulate real-world decision-making scenarios, allowing researchers to isolate variation in judgements that arises from individual differences rather than case-specific factors. In child welfare research, vignette studies are frequently used to examine inconsistencies in professional assessments, identify sources of discretionary variation, and evaluate the impact of background characteristics, training, and contextual influences on judgements. Our study builds on similar investigations conducted in the UK and the US (Lwin et al., 2024; Wilkins & Meindl, 2022) that document a high prevalence of noise in child welfare assessments. However, child welfare systems differ significantly across countries in terms of legal frameworks, professional discretion, and decision-making structures. This study contributes to the literature by being the first to examine statistical noise in a Nordic context, where child welfare services are characterised by a high degree of professional autonomy. Secondly, while previous research has demonstrated high levels of noise, we aim to expand on this by decomposing it into level noise and pattern noise. Level noise is composed of the variability that can be captured by the means of each rater, e.g. if a rater is distinguished by overall leniency towards parents across cases. Pattern noise, on the other hand, arises when raters disagree in less predictable ways: the same rater might be stricter than others on one case, but more lenient on another. This type of variation is a natural consequence of judgements that are complex and discretion-laden, which makes it virtually impossible to identify a single clear cause for why raters diverge. This distinction is important because addressing level noise and pattern noise requires different strategies. Thirdly, does breaking the global assessments down into smaller, domain-specific assessments reduce the noise? This idea builds on findings from other fields (Arkes et al., 2010; Kahneman et al., 2021; Yu & Kuncel, 2020), which suggest that global evaluations – those that cover multiple aspects of an issue – tend to exhibit higher levels of noise compared to disaggregated assessments broken into specific sub-components.

Methods

This study was conducted within a cross-county division of the Norwegian Child Welfare Services. Child welfare workers from a total of 11 counties participated in two surveys. All participants were actively employed as child welfare workers at the time of the study. The majority held a bachelor's degree in child welfare or a related field (94%), while a minority had a master's degree (6%). The first survey was completed by all counties within one month, while the second survey was conducted six months later. Participation in the study was voluntary but encouraged by the heads of the local offices. To ensure consistency, the lead author travelled to each county to distribute paper-based questionnaires. Participants were presented with ten realistic yet fictitious vignettes. We chose vignette studies because they make it possible to gather a representative and varied sample of judgements, while at the same time reflecting real-life decision-making situations where child welfare workers must often act on the basis of limited and sometimes ambiguous information. The two Norwegian authors created the vignettes together with three heads of agency, all of whom have decades of experience from child welfare work. The aim was to ensure realism of the descriptions while retaining anonymity. The vignettes were designed to span a range of severity and encompass various domains of child welfare concerns. For instance, one vignette included a lengthy and detailed description of a caseworker's suspicions of violence and mistreatment. Another vignette consisted of a brief referral from a kindergarten, providing minimal information apart from observations of a child exhibiting worrisome behaviour. All three heads of agencies agreed that the vignettes were realistic, and a fair representation of the varied decisions child welfare workers faced in the field. The same ten vignettes were evaluated by participants in both Survey 1 and Survey 2. All vignettes were originally written in Norwegian. Both the original Norwegian and AI-translated English versions are available online at the Open Science Framework (see https://doi.org/10.17605/OSF.IO/A9NQ2). Survey 1 and 2 had almost no missing data (1.6% and 0.57% of responses respectively). Participants were requested not to discuss their evaluations or the vignette materials with one another until after the completion of the second survey. This project does not have prior ethical approval, as it was originally initiated as an internal development initiative. Participation was entirely voluntary, and no information that could identify respondents was collected.

Survey 1

Participants ($N = 49$) were first asked to provide a global care evaluation for each vignette using an integer scale from 1 to 9, where the endpoints were defined as follows: 1 = Severe neglect, and 9 = Normal care. Next, they were asked to indicate whether they believed that child welfare services should initiate an investigation. Finally, participants were asked to evaluate the potential effectiveness of child welfare interventions in improving the situation, using a scale from 1 to 9, where 1 = Not effective and 9 = Very effective. Our motivation for including the effectiveness of intervention assessment was to ensure that global care evaluations were isolated from participants' beliefs about the effectiveness of potential interventions. For example, in situations of severe neglect, we wanted respondents to clearly distinguish between their assessment of the care situation itself and their belief about whether the situation could be improved through child welfare interventions, allowing these to be captured as two separate and distinct responses.

Survey 2

In the second survey (*N* = 44), the global care evaluation was replaced with more detailed assessments across specific domains of concern, for example in relation to alcohol or drug misuse, poor interpersonal skills, risk of violence or sexual abuse, and parental mental health problems. For each domain, participants rated their level of concern on an integer scale from 1 (Not at all worried) to 9 (Extremely worried) as shown in Figure 1. Each domain was accompanied by a brief descriptor. For example, the domain 'alcohol and drug misuse' was described as: 'Concerns whether caregivers use alcohol or other substances and the extent to which this may impact the child and the caregivers' ability to

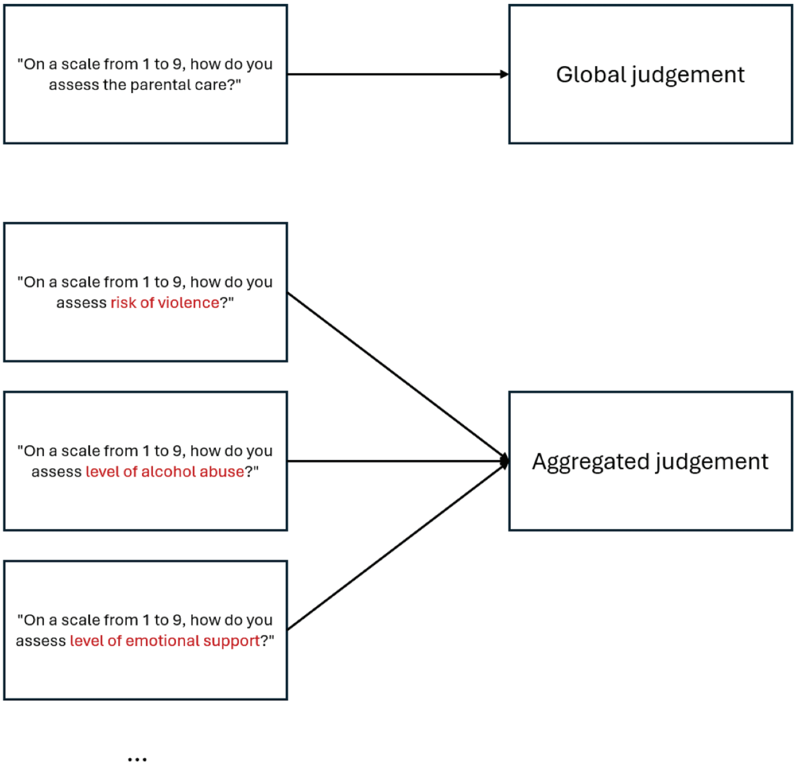
"On a scale from 1 to 9, how do you assess the parental care?"
Global judgement
"On a scale from 1 to 9, how do you assess risk of violence?"
"On a scale from 1 to 9, how do you assess level of alcohol abuse?"
"On a scale from 1 to 9, how do you assess level of emotional support?"
...
Aggregated judgement

Figure 1. Contrasting global and disaggregated evaluation methods.

provide care'. Next, we calculated the average rating for each vignette across multiple domains, rounding the results to the nearest integer. Four vignettes included ratings across four domains, three vignettes had ratings across five domains, and three vignettes had ratings across three domains. The rounded mean of the ratings for each vignette was then used as a proxy for a global care assessment. Both the original Norwegian questionnaires and AI-translated English versions are available online at the Open Science Framework (see https://doi.org/10.17605/OSF.IO/A9NQ2).

Statistical analysis

We started by plotting the data for each vignette using histograms. This provided an overview of the noise represented by the varied assessments, highlighting the spread of ratings across cases. Visualising the data in this way allowed us to identify patterns of variability and assess the extent of disagreement among raters. Next we utilised Gwet AC2 to measure inter-rater reliability (Gwet, 2014). The benefits of this procedure are its robustness to missing data and its flexibility in handling various data structures. Unlike traditional measures such as Cohen's kappa or weighted kappa, which can be overly sensitive to marginal distributions, Gwet's AC2 provides more stable reliability estimates, even when the prevalence or distribution of scores varies among raters. Our data included a continuous scale ranging from 1 to 9, and Gwet's AC2 allowed us to apply different weighting schemes to account for the nature of the ratings. Specifically, we applied linear weights to model the severity of disagreements in the ratings. The ability to flexibly implement ordinal, linear, or quadratic weighting schemes ensures that the measure of reliability aligns with the conceptual importance of rater agreement within our study context. This feature is particularly valuable when the distance between ratings carries a meaningful interpretation, as it does in our scale. There are no universal interpretation standards for Gwet's AC2, as the context of professional judgements is crucial. However, Landis and Koch (1977) provide the following interpretation for Cohen's Kappa, a similar reliability measure: values below 0.00 indicate poor agreement, 0.00–0.20 suggest slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 indicate almost perfect agreement. Given the potentially serious consequences of child welfare decisions, which may include removal of the child from their family home, it can be argued that these services should aim for consistently high reliability rates.

To estimate the degree of level noise in our dataset, we followed the approach outlined by Kahneman et al. (2021). Level noise represents the variability in judgements attributable to systematic differences in the mean ratings given by individual raters, regardless of the specific vignettes being evaluated. Total noise was calculated as the average variance of ratings across vignettes, capturing the overall variability in the dataset. To compute level noise, we calculated the mean rating for each rater across all vignettes and calculated the variance of these mean ratings. This variance represents the level noise, which was expressed as a proportion of the total noise by dividing the variance of the rater means by the total noise. This procedure is equivalent to performing an analysis of variance (ANOVA) to quantify the proportion of variability attributable to raters. Rater and case effects are fixed factors, and the sum of squares associated with raters, reflect the variability in their mean judgements. This value can then be divided by the sum total of the rater sums of squares and the residual sum of squares. Both methods produced consistent estimates of the proportion of variability explained by level noise. The data and all analysis scripts for this study are available at the Open Science Framework (OSF). See https://doi.org/10.17605/OSF.IO/A9NQ2.

Results

We found that the care assessments had a substantial degree of noise, with large differences between vignettes. As shown in Figure 2, some of the vignettes are difficult to distinguish from a random selection on a nine-item scale. The average ratings, across vignettes, from Experiment 1 and 2 were also similar (5.78 and 5.43 respectively). We also found that the proportion of level noise was high with 24.7% in Experiment 1 and 28.7% in Experiment 2.

The inter-rater reliability was .501 for Experiment 1 (95% CI [.399, .603]) and .543 for Experiment 2 (95% CI [.466, .620]). To statistically compare the relative effectiveness of domain-specific assessments (Experiment 2) with global assessments (Experiment 1) on

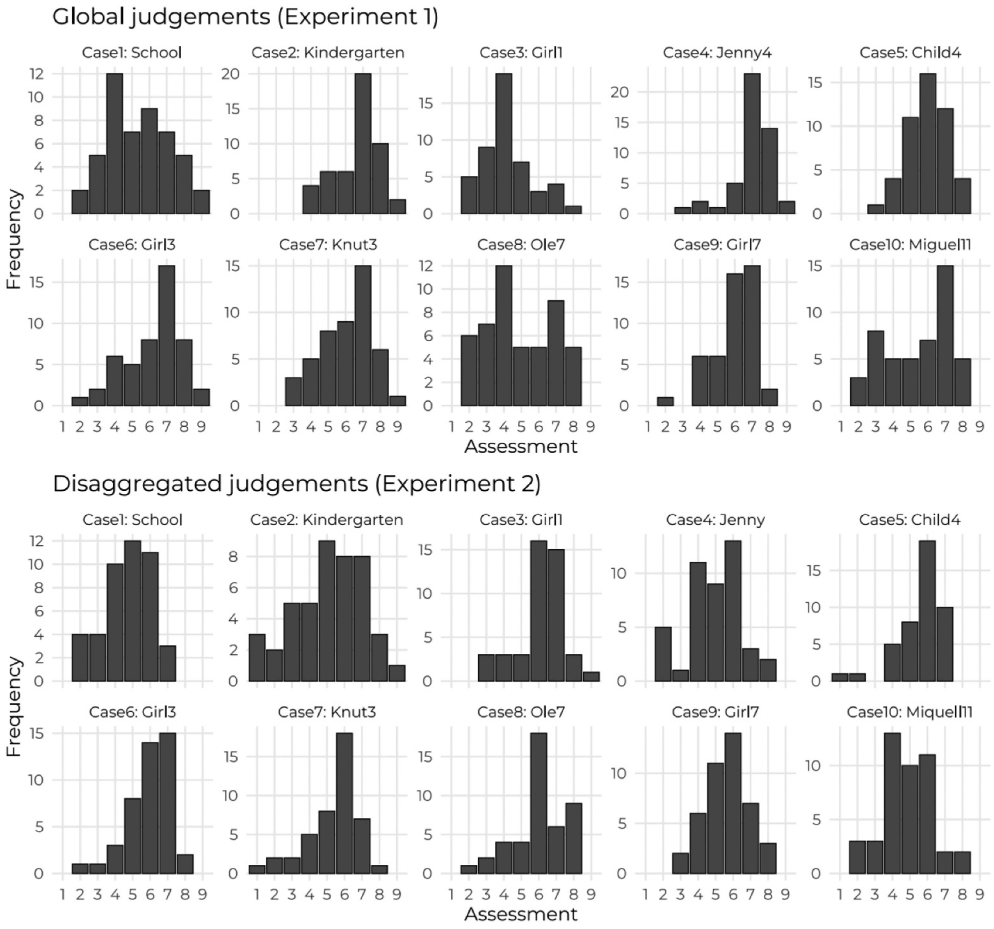

Figure 2. Histograms of judgements.

inter-rater reliability, we conducted a bootstrap analysis. Random sampling with replacement yielded a 95% confidence interval of −0.05 to 0.13, indicating no significant improvement.

Discussion

Variability in child welfare professional judgements is well-documented, but this study is the first to specifically quantify the extent of noisy evaluations in child welfare service assessment. There is strong justification to believe that while some variation is inevitable (Munro, 2019) – and even beneficial in certain contexts – noisy variation represents an undesirable inconsistency that undermines the quality and fairness of professional judgements. Recognising that absolute truth is often unknowable, it is particularly important to focus on noise, which can be objectively measured and reduced. Our findings also revealed that a substantial portion of the noise stemmed from level noise – the variability in judgements driven by systematic differences in the mean ratings of individual raters, independent of the specific vignettes being evaluated. This is important because level noise is more amenable to reduction compared to stochastic, unidentified noise. To mitigate level noise, workers could compare their independent ratings with those of their peers to better understand their own rating tendencies relative to others. This strategy has been implemented in other settings such as police decision making (Wire et al., 2024). The process usually involves providing anonymised comparative feedback (e.g. 'you rate lower than 80% of peers on average'), followed by structured opportunities to reflect on these differences. Such feedback does not prescribe a 'correct' answer but highlights systematic tendencies that might otherwise go unnoticed. While this approach directly targets level noise by making individual thresholds more transparent, it does not eliminate other sources of variability, such as pattern noise or broader systemic differences. Conducting a noise audit is relatively straightforward, requiring only a series of independent quantitative assessments of a vignette or case file. The greater challenge lies in convincing stakeholders of the value of quantifying noise and how to manage subsequent feedback, as this necessitates systemic changes that are often difficult to implement as well as potentially challenging professional self-perceptions and indeed deciding whether the workers who are 'more lenient' or 'less lenient' are the ones who need to change (Munro, 2019). Moreover, research has shown that workers' underlying attitudes towards issues such as removal, reunification, and parental participation can significantly shape their judgements, meaning that noise is not only a matter of calibration but also of deeply held professional values (Davidson-Arad & Benbenishty, 2014).

In addition to noise audits, standardised procedures and assessments provide a promising approach to reducing noise. As Turney et al. (2024) argues, there is a need for clarity and transparency around decision thresholds, for example the line between waiting to observe how a situation develops and initiating the forcible removal of a child. However, efforts to implement standardised practices in child welfare judgements often encounter substantial challenges and resistance from practitioners – a response that is understandable given the importance of professional discretion in this field. However, other professions, such as psychiatry and criminal justice, have successfully implemented more standardised approaches (Berk et al., 2021; Meehan et al., 2022). This raises the question of whether there are lessons to be learned from these fields – not necessarily about the mechanics of creating standardisation, but about the cultural strategies for introducing it in ways that align with practitioners' professional values and goals. Such approaches could help practitioners view standardisation not as a constraint, but as a tool to enhance their ability to achieve positive outcomes. In the field of employee selection, Nolan and Highhouse (2014) demonstrate experimentally that the need for autonomy is a driving force behind resistance against standardisation, despite improving decision-making. There is an inherent imbalance in standardisation efforts: noise often goes unnoticed, making its negative impact less visible, while attempts to reduce noise, such as implementing guidelines, training programmes, or peer comparisons, are highly conspicuous. In many ways, this reflects the situation in social work practice, in which the harms caused by errors of non-intervention are often tragic and highly salient, compared with the more diffuse and overlooked harms caused by errors of over-intervention. As a result of this imbalance, potentially valuable interventions to improve the quality of decision-making may be erroneously rejected. Addressing the imbalance requires transparent communication about the presence and consequences of noise, alongside collaborative approaches to implementing noise-reduction strategies that respect and incorporate the perspectives of practitioners.

We did not find evidence that subdividing global decisions into sub-assessment reduces noise, which was unexpected given the positive findings from other fields (Arkes et al., 2010; Yu & Kuncel, 2020). Disaggregating global, undefined decisions into clearly operationalised sub-assessments is widely considered a gold standard for noise reduction (Kahneman et al., 2021). Our findings may reflect the unrefined nature of the sub-assessments. For instance, asking respondents to rate the level of concern regarding alcohol use requires them to define what constitutes concern, which may be no easier than providing a global assessment of care capacity. Despite these results, carefully breaking down and operationalising decisions remains a prudent approach to reducing noise that remains worthy of further study. Future research should focus on refining these methods to provide decision-makers with tools that effectively minimise noise. Perhaps the most effective way to reduce noise, however, lies in aggregating independent peer judgements, as exemplified by the Delphi process and similar structured consensus methods. Future research should focus on refining these methods to provide decision-makers with tools that effectively minimise noise.

At the same time, most studies on noise have concentrated on identifying the source of noise (Fluke et al., 2016). While this line of research is valuable, it may have limits if a substantial part of the variation is essentially stochastic, arising from the discretion-laden nature of many child welfare judgements. For example, when a report of concern arrives almost devoid of content, there is virtually no basis for judgement. Any decision will be noisy unless sensible, evidence-based rules or decision aids guide the process. In our view, more attention should be directed towards reducing this residual, 'unidentified' noise through strategies such as operationalising criteria, aggregating independent judgements, and, more broadly, limiting the extent to which outcomes hinge on individual discretion. Since child welfare services are most often organised at the municipal level, different offices may develop their own practices for handling situations with little information. This can reduce consistency, as outcomes may depend on noisy local routines rather than a shared, well-developed instructions that could provide greater uniformity across the system.

Conclusion

In summary, this study demonstrates a significant amount of noise in Norwegian child welfare assessments. A substantial portion was level noise, which can be addressed through feedback and correction, while a considerable share was pattern noise, suggesting the need for standardisation routines. Notably, breaking down global decisions into sub-assessments did not reduce noise, highlighting the importance of refining and operationalising these approaches. Lastly, we want to emphasise that the results we present do not reflect the typical decision-making process within child welfare services in Norway. In practice, major decisions are made collaboratively, with a larger team thoroughly reviewing all relevant data from the case file.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Lene Caroline Ljosland Nordmo is a clinical psychologist working in the Norwegian child welfare services. She specialises in supporting children and families facing complex challenges, with a focus on promoting resilience and informed decision-making. Drawing on her clinical experience, Nordmo is particularly interested in how psychological knowledge can enhance assessments, interventions, and professional practice within child welfare settings.

David Wilkins is a researcher and senior lecturer at Cardiff University. His work focuses on decision-making in child welfare services, with particular emphasis on supervision, professional judgement, and practice improvement. Wilkins has published widely on how social workers assess risk and make decisions, and he is committed to bridging research and practice to support better outcomes for children and families.

Magnus Nordmo is a researcher at the University of South-Eastern Norway (USN). His work focuses on decision-making, mental health, and child welfare services. With a background in psychology, Nordmo explores how professional judgements are shaped by individual, organisational, and societal factors. His current research examines the intersection of developmental psychology, behavioural genetics, and child welfare practice.

References

Arkes, H. R., González-Vallejo, C., Bonham, A. J., Kung, Y.-H., & Bailey, N. (2010). Assessing the merits and faults of holistic and disaggregated judgments. *Journal of Behavioral Decision Making*, *23*(3), 250–270. https://doi.org/10.1002/bdm.655

Austin, W., & Thomas, A. (1977). A survey of judges' responses to simulated legal cases: Research note on sentencing disparity. *Journal of Criminal Law & Criminology*, *68*(2), 306. https://doi.org/10.2307/1142852

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 and Neglect*, *49*, 63–75. https://doi.org/10.1016/j.chiabu.2015.03.015

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Burridge, A., & Gill, N. (2017). Conveyor-belt justice: Precarity, access to justice, and uneven geographies of legal aid in UK asylum appeals. *Antipode*, *49*(1), 23–42. https://doi.org/10.1111/anti.12258

Davidson-Arad, B., & Benbenishty, R. (2014). Child welfare attitudes, risk assessments and intervention recommendations: The role of professional expertise. *British Journal of Social Work*, *46*(1), 186–203. https://doi.org/10.1093/bjsw/bcu110

Fluke, J. D., Corwin, T. W., Hollinshead, D. M., & Maher, E. J. (2016). Family preservation or child safety? Associations between child welfare workers' experience, position, and perspectives. *Children & Youth Services Review*, *69*, 210–218. https://doi.org/10.1016/j.childyouth.2016.08.012

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

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*.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown. https://books.google.no/books?id=fhIBEAAAQBAJ

Keddell, E. (2017a). Comparing risk-averse and risk-friendly practitioners in child welfare decision-making: A mixed methods study. *Journal of Social Work Practice*, *31*(4), 411–429. https://doi.org/10.1080/02650533.2017.1394822

Keddell, E. (2017b). Interpreting children's best interests: Needs, attachment and decision-making. *Journal of Social Work*, *17*(3), 324–342. https://doi.org/10.1177/1468017316644694

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lwin, K., & Beltrano, N. (2022). Rethinking evidence-based and evidence-informed practice: A call for evidence-informed decision making in social work education and child welfare practice. *Social Work Education*, *41*(2), 166–174. https://doi.org/10.1080/02615479.2020.1819973

Lwin, K., Hoagland, A., Antwi-Boasiako, K., MacKenzie, P., & Fallon, B. (2024). Examining the role of child welfare worker characteristics and the substantiation decision. *Child Abuse and Neglect*, *149*, 106641. https://doi.org/10.1016/j.chiabu.2024.106641

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry*, *27*(6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4

Meindl, M., & Wilkins, D. (2021). Can social workers forecast future actions, events, and outcomes? A study of referrals to children's services in England. *Child Care in Practice*, *31*(2), 321–335. https://doi.org/10.1080/13575279.2021.2001434

Mosteiro, A., Beloki, U., Sobremonte, E., & Rodríguez, A. (2018). Dimensions for argument and variability in child protection decision-making. *Journal of Social Work Practice*, *32*(2), 169–187. https://doi.org/10.1080/02650533.2018.1439459

Munro, E. (1999). Common errors of reasoning in child protection work. *Child Abuse and Neglect*, *23*(8), 745–758. https://doi.org/10.1016/S0145-2134(99)00053-8

Munro, E. (2019). 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

Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance*, *27*(4), 328–346. https://doi.org/10.1080/08959285.2014.929691

Rustad, K. B., Lauritzen, C., Skaale Havnen, K. J., Fossum, S., Christiansen, Ø., & Vis, S. A. (2022). The impact of case factors on the initial screening decision in child welfare investigations in Norway. *Child Abuse and Neglect*, *131*, 105708. https://doi.org/10.1016/j.chiabu.2022.105708

Taylor, B. J. (2012). Models for professional judgement in social work. *European Journal of Social Work*, *15*(4), 546–562. https://doi.org/10.1080/13691457.2012.702310

Taylor, B., & Whittaker, A. (2018). Professional judgement and decision-making in social work. *Journal of Social Work Practice*, *32*(2), 105–109. https://doi.org/10.1080/02650533.2018.1462780

Turney, D., Alfandari, R., Taylor, B. J., Ghanem, C., Helm, D., Killick, C., Lyons, O., O'Leary, D., Ebsen, F., & Bertotti, T. (2024). Threshold decisions in social work: Using theory to support practice. *British Journal of Social Work*, *54*(7), 2996–3013. https://doi.org/10.1093/bjsw/bcae073

Wilkins, D., & Meindl, M. (2021). *A randomised controlled trial of a 'checklist' intervention to mitigate confirmation bias and improve forecasting accuracy in social work*. https://orca.cardiff.ac.uk/id/eprint/143804/

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*, *31*(2), 301–320. https://doi.org/10.1080/13575279.2022.2118674

Wilkins, D., & Meindl, M. (2023). Can social workers estimate the likelihood of future actions and events? A forecasting accuracy study. *British Journal of Social Work*, *54*(3), 1150–1169. https://doi.org/10.1093/bjsw/bcad234

Wilkins, D., & Meindl, M. (2024). Measuring the ratio of true-positive to false-positive judgements made by child and family social workers in England: A case vignette study. *Child & Family Social Work*, *29*(2), 327–338. https://doi.org/10.1111/cfs.13086

Wire, S., Mitchell, R. J., & Schiess, J. (2024). Consistently inconsistent: Examining variability in police decision-making in mental health calls using a novel noise audit approach. *Journal of Criminal Justice*, *93*, 102201. https://doi.org/10.1016/j.jcrimjus.2024.102201

Yu, M., & Kuncel, N. (2020). Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions*, *6*(2). https://doi.org/10.25035/pad.2020.02.002