Development of the ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS):

TECHNICAL REPORT

ANROWS

ANROWS acknowledgement

This material was produced with funding from the Australian Government Department of Social Services. Australia's National Research Organisation for Women's Safety (ANROWS) gratefully acknowledges the financial and other support it has received from the government, without which this work would not have been possible. The findings and views reported in this paper are those of the authors and cannot be attributed to the Australian Government Department of Social Services.

Acknowledgement of Country

ANROWS acknowledges the Traditional Owners of the land across Australia on which we live and work. We pay our respects to Aboriginal and Torres Strait Islander Elders past and present. We value Aboriginal and Torres Strait Islander histories, cultures and knowledge. We are committed to standing and working with First Nations peoples, honouring the truths set out in the Warawarni-gu Guma Statement.

Acknowledgement of lived experiences of violence

ANROWS acknowledges the lives and experiences of the women and children affected by domestic, family and sexual violence who are represented in this report. We recognise the individual stories of courage, hope and resilience that form the basis of ANROWS research.

Caution: Some people may find parts of this content confronting or distressing. Recommended support services include 1800RESPECT (1800 737 732), Lifeline (13 11 14) and, for Aboriginal and Torres Strait Islander people, 13YARN (13 92 76).

A catalogue record for this
book is available from the
National Library of Australia

Published by

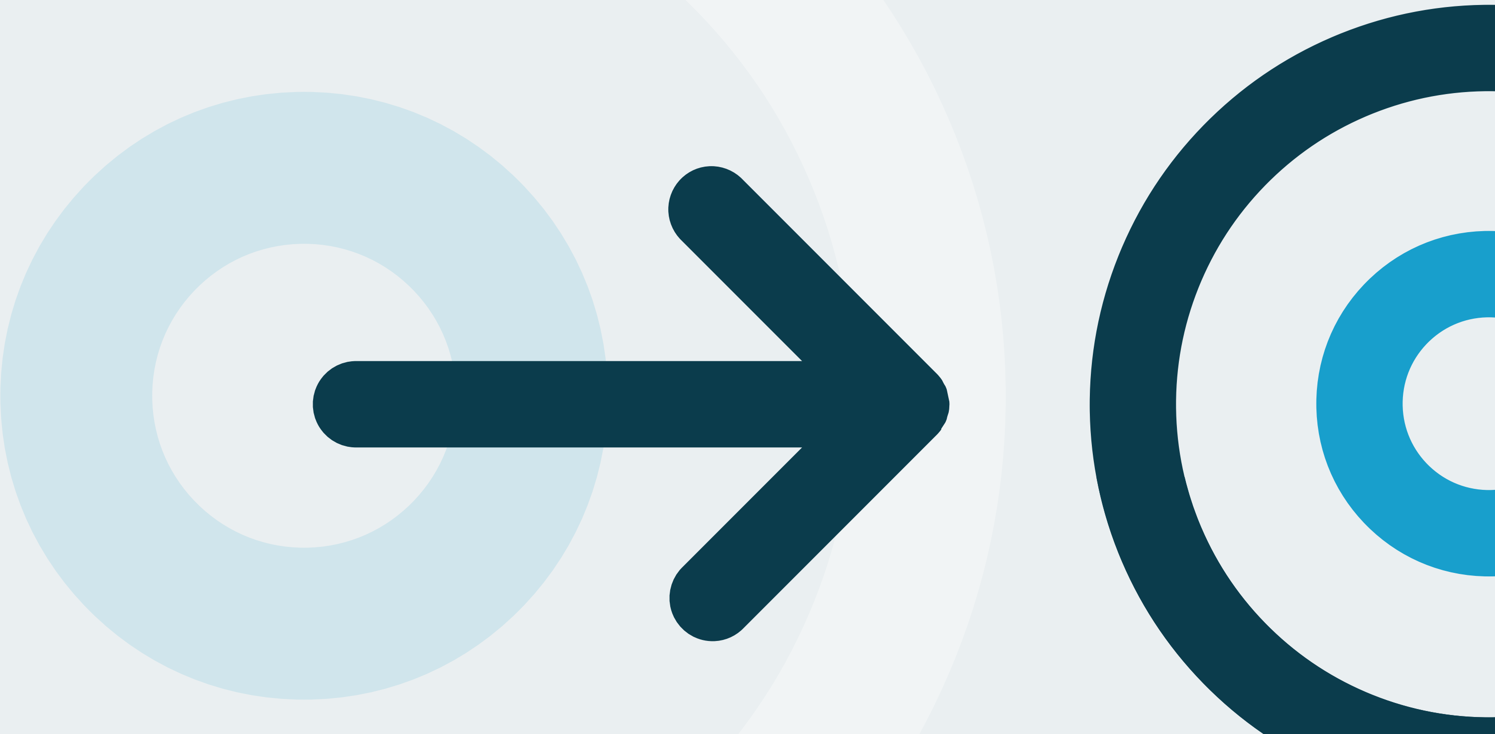
Australia's National Research Organisation for Women's Safety Limited (ANROWS)
PO Box Q389, Queen Victoria Building, NSW 1230 | www.anrows.org.au
ABN 67 162 349 171

Development of the ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS):

TECHNICAL REPORT

Prepared by

Lorelei Hine, Research Manager (Acting), ANROWS
Angela Higginson, Associate Professor, Queensland University of Technology
Elizabeth Eggins, Research Fellow, Griffith University
Charlotte Bell, Senior Research Officer, ANROWS
Elizabeth Watt, Research Manager, ANROWS



Author acknowledgement

The authors would like to thank the Evidence Portal Policy Reference Group and the Evidence Portal Advisory Group for their comments and input on the development of the risk of bias tool. We would like to also thank Professor David B. Wilson for his invaluable comments and advice. Thanks to the following ANROWS staff: Dr Jane Lloyd for her research expertise and Dr Jennifer Sijnja for her knowledge translation expertise and support. Lastly, we would like to thank the following ANROWS staff members who were integral to the early testing of the tool: Dr Brigitte Gilbert, Liuissa Zhen, Sara Roetman, Dr Rebecca Pollard and Rhiannon Smith.

Suggested citation

Hine, L., Higginson, A., Eggins, E., Bell, C., & Watt, E. (2023). *Development of the ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS): Technical report* (Evidence Portal technical report, EP.23.2/1). ANROWS.

Contents

Executive summary

The Australia's National Research Organisation for Women's Safety (ANROWS) Evidence Portal of interventions to address and end violence against women (the Evidence Portal) is a living resource that provides policymakers and practitioners with access to evidence on the nature and effectiveness of interventions designed to address and end violence against women in Australia and other high-income countries. The Evidence Portal is a living online resource and can be accessed here: **https://www.evidenceportal.au**

The ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS) is a bespoke risk of bias tool developed as part of the Evidence Portal. It has been designed for use with the quantitative impact evaluations included in the Evidence Portal as well as for systematic reviews in the social and psychological sciences more broadly. The tool is designed to be applied to quantitative impact evaluations of interventions to critically appraise them across six domains. Collectively, these domains examine whether the design, reporting and implementation of an evaluation study can support the conclusion that the intervention caused a change in the measured outcomes, or if study flaws are likely to lead to over- or underestimates of the effect of the intervention.

This report briefly introduces risk of bias tools and sets out the justification for creating a bespoke risk of bias tool for use in the Evidence Portal. It details the development of the ANROWS-IRIS using a five-stage adaptive approach, incorporating: 1) scanning; 2) identification; 3) design; 4) evaluation; and 5) adaptation.

The resultant ANROWS-IRIS assesses risk of bias of studies across six domains:

1) study design
2) selection bias
3) confounders
4) data collection methods
5) withdrawals and drop-outs
6) intervention integrity and fidelity.

This report describes the importance of each domain, as well as the approach to generating an overall risk of bias rating for a study across the domains. The overall risk of bias rating allows for a simple, user-friendly evaluation of the methodological rigour of studies. Initial interrater reliability testing is reported for a sample of 11 studies from the Evidence Portal, with results showing good agreement between assessors across each domain and overall.

The Evidence Portal is strengthened by the ability to assess the risk of bias of included interventions in the field. This lends the findings from the Evidence Portal greater robustness and aims to promote evidence-informed policy- and decision-making that can account for the quality of the literature and confidence in causal findings. The ANROWS-IRIS is a valuable research tool to support the work of the Evidence Portal.

The guidance document for assessors and the full ANROWS-IRIS tool, including the rating rubrics for the domains and overall ratings, are published separately (see Higginson et al., 2023 and Eggins et al., 2023 respectively).

Introduction

The ANROWS Evidence Portal of interventions to address and end violence against women (hereafter referred to as the Evidence Portal) is a living resource that provides policymakers and practitioners with access to evidence on the nature and effectiveness of interventions designed to address and end violence against women in Australia and other high-income countries. While the public-facing Evidence Portal website is designed to provide plain-language information on evaluations of interventions that are easily understood by a variety of audiences, the project is underpinned by a rigorous systematic search and screening methodology that is informed by best practice standards for systematic reviews (Higgins et al., 2022; Moher et al., 2009; The Campbell Collaboration, 2021).

A key component of evidence-based research and practice in this space is appraising the credibility of conclusions drawn from studies of intervention effectiveness. In the context of violence against women, the Evidence Portal aims to confidently answer the question, "Does the intervention in question 'work' to prevent, identify, respond to, or promote recovery and healing from violence against women?" This requires a critical appraisal of quantitative evaluation studies across multiple domains that collectively examine whether the design, reporting and implementation of an evaluation study can support the conclusion that the intervention caused a change in the measured outcomes, or if study flaws are likely to lead to over- or underestimates of the effect of the intervention.

This approach to critically appraising evaluation research is called risk of bias assessment. Risk of bias assessment typically requires an assessor to answer a series of signalling questions to identify and rate the impact of several potential sources of bias. These signalling questions are often grouped into domains or categories of potential threats to study credibility (e.g. bias resulting from participant selection, data collection or reporting). The answers to these signalling questions are then scored to assign ratings (e.g. high, medium or low risk of bias) to each domain of potential bias and, in some cases, to the study overall.

There are many existing tools for assessing risk of bias; however, these vary widely by the types of studies appraised, the level of ambiguity in tool guidance, the consistency in how they are applied, and the contextual suitability to specific fields of study (Drukker et al., 2021; Jüni et al., 1999; Page et al., 2018; Quigley et al., 2019; Seehra et al., 2016; Waddington et al., 2017). Although several risk of bias tools are able to evaluate a wide range of study designs (see for example the Effective Public Health Practice Project [EPHPP] Quality Assessment Tool for Quantitative Studies [EPHPP, 2009]; Risk Of Bias In Non-randomized Studies of Interventions [ROBINS-I; Sterne et al., 2016a]; and the Joanna Briggs Institute (JBI) checklists [JBI, 2020]), none met all of the needs of the Evidence Portal.

To function most effectively, the Evidence Portal required a brief, intuitive yet comprehensive tool that could be applied by trained non-experts and understood by a variety of audiences. The tool needed to:

- evaluate quantitative primary studies across a wide range of study designs, from pre–post single group studies through to randomised controlled trials
- assess non-randomised studies as providing causal evidence with a low risk of bias if such studies are conducted with a strong focus on internal validity
- recognise a more nuanced spectrum of risk of bias and differentiate between studies with a finer granularity than simply high, medium or low risk of bias.

Such a tool would not only satisfy the requirements of the Evidence Portal but would have direct applicability to systematic reviews in the social sciences and violence against women research more broadly. This report details the development of ANROWS's risk of bias tool – the ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS). This includes the process taken to determine the domains, the early testing and refinement, the final domains of the tool and interrater reliability testing.

Development of the ANROWS-IRIS

The ANROWS-IRIS was developed using a five-stage adaptive approach, incorporating: 1) scanning; 2) identification; 3) design; 4) evaluation; and 5) adaptation.

Scanning

The project team began by reviewing the extant quality assessment literature, including methodological sources and existing tools. The following tools were evaluated in the initial scanning phase:

- Checklist for measuring study quality (Downs & Black, 1998)
- Revised Cochrane risk of bias tool for randomized studies (RoB 2; Higgins et al., 2016)
- EPHPP Quality Assessment Tool for Quantitative Studies (EPHPP, 2009, 2010; Thomas et al., 2004)
- Cochrane Effective Practice and Organisation of Care (EPOC) resources for review authors (EPOC, 2017)
- JBI checklist for quasi-experimental studies (JBI, 2020)
- The Observational Study Quality Evaluation (Drukker et al., 2021)
- Risk of Bias Assessment Tool for Nonrandomized Studies (RoBANS; Kim et al., 2013)
- National Institute for Health and Clinical Excellence (NICE) quality appraisal checklist – quantitative intervention studies (NICE, 2012)
- ROBINS-I assessment tool (Sterne et al., 2016a, 2016b)
- Scottish Intercollegiate Guidelines Network (SIGN) 50 (SIGN, 2019)
- Critical appraisal tool (White et al., 2020).

The project team also consulted academic literature on risk of bias and study quality assessments, including key works by Alexander et al. (2015), Berger (2005), Higgins et al. (2011), Higgins et al. (2022), Jüni et al. (1999), Quigley et al. (2019) and Waddington et al. (2017).

Identification

The project team then collated the common domains across tools and the way each had been appraised and examined, discussed and consulted on the suitability of each tool based on ANROWS's specifications and coverage of the key risk of bias domains. Each tool was assessed for its potential requirements across staffing skill, time and resources, as well as its utility for use with the extant violence against women evidence. Some tools did not meet the specifications because they had lengthy assessment times (e.g. Cochrane RoB 2, ROBINS-I), because they had subjectivity in ratings (e.g. JBI), or because they did not cover all study designs (e.g. RoB 2) or had separate scales for each study design (e.g. JBI). The identification stage led to a list of core risk of bias domains to be included in the ANROWS tool.

Design

The first draft of the tool was framed around study confidence rather than risk of bias. The project team designed a draft tool, incorporating 13 signal items across six domains, culminating in a five-point overall rating scale: very high, high, moderate, low or very low confidence.

Evaluation

An iterative process of evaluation was conducted. The initial draft was distributed to the Evidence Portal Advisory Group for their feedback. This feedback was integrated into the tool which was tested by the project team.

Testing was conducted on sets of studies that evaluated specific interventions using a diverse range of study designs (e.g. cognitive behavioural therapy for victims and survivors of violence against women with post-traumatic stress disorder [PTSD], second-responder policing interventions). Testing focused on assessing the consistency of ratings across assessors, evaluating the ease or difficulty of applying the tool and identifying areas where more guidance was needed, comparing study ratings to those in published systematic reviews and assessing the face validity of the final rating. We also sought expert feedback from Professor David B. Wilson, a leading scholar in the field of evidence synthesis in the social sciences.

Adaptation

Feedback from testing and expert consultation was incorporated into the draft of the tool. The final version of the ANROWS-IRIS includes 19 signal items across six domains, culminating in a six-point overall rating scale: very low, low, moderate, moderate-high, high or very high risk of bias. The shift from the original framing as a "study confidence" tool to a "risk of bias" tool retains consistency with other tools that tend to use terminology around risk of bias.

ANROWS-IRIS risk of bias domains

The ANROWS-IRIS assesses risk of bias across six domains:

1) study design; 2) selection bias; 3) confounders; 4) data collection methods; 5) withdrawals and drop-outs; and 6) intervention integrity.

The guidance document for assessors and the full tool, including the rating rubrics for the domains and overall ratings, are published separately (Higginson et al., 2023; Eggins et al., 2023).

Domain 1: Study design

Domain 1 contains two signal items:

Q1 Select the study design.

Q2 Is the comparison condition or group comprised of treatment refusers or drop-outs?

Domain 1 provides an initial assessment of the study's risk of bias based on the study design. The subsequent domains are then used to upgrade or downgrade the study's overall risk of bias. Studies are categorised as either randomised controlled trials, quasi-experimental impact evaluations with comparison group(s), long interrupted time-series designs without comparison group(s) or single group pre–post designs.

Certain study designs are inherently more limited than others in their ability to create causal evidence. Study design is often conceptualised as a hierarchy, with randomised controlled trials towards the top of the "evidence pyramid" and single group pre–post designs towards the bottom. The study design category also creates a skip structure in the tool because certain domains are only applicable to studies that include a comparison group (Domain 3: Confounders and Domain 5: Withdrawals and drop-outs).

Studies that use a comparison group comprised of treatment refusers or drop-outs are unable to provide causal evidence because this element of the study design introduces a critical risk of bias (Sterne et al., 2016b). In this instance, the assessor would score Domain 1 and finish the risk of bias assessment and the study would be rated as having a very high risk of bias. Domains 2 to 6 are only completed for studies where there is either no comparison group or the comparison group is no treatment, a waitlist control, placebo, treatment/business as usual, an alternative treatment or an alternative dosage of the treatment.

Domain 2: Selection bias

Domain 2 contains six signal items:

Q3 Do the authors clearly describe the target population?

Q4 Do the authors clearly describe the sampling frame?

Q5 Is the sampling frame likely to be appropriate for the target population?

Q6 Do the authors clearly describe the sampling approach?

Q7 Are the study participants likely to be representative of the sampling frame?

Q8 Do the authors demonstrate that the participants are likely to be representative of the target population?

Domain 2 takes a hybrid approach to the assessment of selection bias and includes items that address both the external and internal validity of the study. Selection bias is concerned with the representativeness of the sample to the target population (Alexander et al., 2015). Selection bias may result in an under- or overestimate of the “true” effect of the intervention when applied to the intervention’s target population. The assessment of selection bias involves an interplay between the aims of the intervention, the possible population of eligible participants and the sampling approach of the study. The questions in this domain focus on participants’ initial selection into the study and not on their allocation to intervention or comparison group(s).

Studies that can demonstrate that their sample is representative of their target population are more likely to produce results that can be generalised to real-world implementations. Domain 2 is rated as a low risk of bias if the study either directly demonstrates that the participants are likely to be representative of the target population or the assessor assesses that the sampling frame is likely to be appropriate for the target population and the study participants are likely to be representative of the sampling frame.

Domain 3: Confounders

Domain 3 contains three signal items and is not scored for single group pre–post study designs:

Q9 Do the authors state or demonstrate if the comparison group was equivalent to the treatment group prior to the intervention?

Q10 Are there any meaningful differences between the groups?

Q11 Do the authors attempt to control for confounding factors in their analysis?

Domain 3 considers whether the causal attributions and estimates of effect are likely to be impacted by confounding factors. Confounding can occur when observable or non-observable factors (other than the intervention) influence the outcome of the study (Waddington et al., 2017). Examples of possible confounders relevant to violence against women studies include age, ethnicity, gender, sexuality, education, marital status, family structure, socio-economic status (e.g. income, class), health status, prior contact with the criminal justice system, prior victimisation, pre-intervention score on outcome measure(s) and/or seasonality or autocorrelation in time-series designs. This is not an exhaustive list, as confounders tend to be contextually dependent.

Domain 4: Data collection methods

Domain 4 contains two signal items:

Q12 Do the outcomes have face validity?

Q13 Do the authors describe how they measured each outcome?

Domain 4 assesses the impact of the methods used to measure the outcomes of the evaluation. When outcome measurements are valid and reliable, the risk of bias is reduced (Thomas et al., 2004). Validity can include face validity (EPHPP, 2009), which is the extent to which a test appears (at face value) to measure what it purports to measure. Face validity can be a simpler assessment for a trained non-expert to make than other forms of validity. This domain also assesses reliability by evaluating the degree to which all outcomes are clearly named and described in replicable detail or whether the authors provide a citation to an existing standardised measure.

Domain 5: Withdrawals and drop-outs

Domain 5 contains three signal items and is not scored for single group pre–post study designs:

Q14 Is there a meaningful difference in attrition or drop-out between the treatment and comparison group?

Q15 Is the attrition systematic or at random?

Q16 If systematic, did the authors control for the impact of differential attrition?

Domain 5 assesses the degree of differential attrition in the study. Differential attrition can introduce bias if the effect of the intervention is an under- or overestimate of the "true" effect that would be obtained if no attrition occurred (Sterne et al., 2016b). The risk of bias due to differential attrition is higher when the attrition is systematic and not controlled for in the analyses (Sterne et al., 2016b). Attrition may occur while the intervention is ongoing, or it may occur after the intervention has been completed but before all outcome measurements are taken. This domain allows the assessor to consider all possible sources of attrition that are reported in the study.

Domain 6: Intervention integrity and fidelity

Domain 6 contains three signal items:

Q17 Was the intervention implemented as intended (as per protocol)?

Q18 Did the authors report that co-intervention or contamination occurred?

Q19 If contamination or co-intervention was reported, did the authors report the results of relevant sensitivity analyses?

Domain 6 assesses the impact of intervention integrity by evaluating whether the intervention was implemented as intended and, further, if co-intervention or contamination occurred. Co-intervention is where participants in the intervention group receive an additional intervention beyond what was intended. This may range from an additional intervention component through to an entirely separate additional intervention. Contamination is when the comparison group receives some or all of the intervention. Deviations from the intended intervention can result in an under- or overestimate of the effectiveness of an intervention (Sterne et al., 2016b), although in some cases this bias may be accounted for in the analysis (Waddington et al., 2017).

Overall risk of bias rating

Risk of bias tools vary as to whether they provide only domain-specific ratings or whether they allow for an overall risk of bias rating. The ANROWS-IRIS provides an overall risk of bias rating to allow a simple, user-friendly evaluation of the methodological rigour of the research in the Evidence Portal. Table 1 shows how the ratings for each of Domains 1 to 6 are combined to give an overall rating of very low, low, moderate, moderate-high, high or very high risk of bias. This six-point rating scale enables a high degree of differentiation between studies.

Table 1: Overall risk of bias rating

Overall risk of bias	Domain 1	Domains 2 to 6	Selection
Very low	Low	Low on all Domains 2 to 6	○
Low	Low	1 Moderate, 0 High	○
	Moderate	Low on all Domains 2 to 6	○
Moderate	Low	2 or more Moderate, 0 High	○
	Moderate	1 or more Moderate, 0 High	○
Moderate-high	Low OR Moderate	1 or 2 High	○
High	Low OR Moderate	3 High	○
	High	**Not** High on Domains 2, 4 **AND** 6	○
Very high	Low OR Moderate	4 or more High	○
	High	High on any of Domains 2, 4 **OR** 6	○
	Critically high	Any combination	○

Each study begins with a rating on Domain 1 (Study design) which can then be downgraded depending on the ratings for subsequent domains. Any study that uses treatment refusers or drop-outs as a comparison group (Q2) is rated as very high risk of bias. Upgrading can occur if a study is rated as low or moderate on Domain 1 and low across each of the Domains 2 to 6:

- Randomised controlled trials (Domain 1 = low) can be upgraded to very low risk of bias if they are rated as low on every subsequent domain.

- Quasi-experimental designs (Domain 1 = moderate) can be upgraded to low risk of bias if they are rated as low on all subsequent domains.

A study can only be rated as having a very low, low or moderate risk of bias if no individual domain is rated as having a high risk of bias. Studies that include any domain ratings of high risk of bias will be rated as having a moderate-high, high or very high risk of bias, depending on the number of domains affected.

Interrater reliability

Interrater reliability testing is an important component in developing risk of bias tools. It is done to ensure that there is consistency in the application of the tools and in interpretation between assessors and across different studies (Hartling et al., 2012). Interrater reliability testing for the ANROWS-IRIS is ongoing and the tool will continue to be tested. At the time of writing, the project team has undertaken a preliminary interrater reliability activity. Two assessors from the project team who were not closely associated with the tool's development were trained to use the finalised tool using the guidance document (published separately) and during team meetings. The two assessors independently rated the same 11 studies (Aupperle et al., 2013; Beck et al., 2016; Echeburúa et al., 2014; Ferrari et al., 2018; Johnson et al., 2011, 2016; Littleton et al., 2016; Littleton & Grills 2019; Moreira et al., 2022; Smith et al., 2015; Wells, et al., 2019). These studies were chosen from the eligible studies within the Evidence Portal as they shared similarities (e.g. they all evaluated cognitive behavioural therapy for victims and survivors who had experienced violence against women and trauma/PTSD) but also differences (i.e. a range of study designs).

We calculated the interrater reliability using both Cohen's Kappa and weighted Cohen's Kappa (Cohen, 1960, 1968; Glen, n.d). Once independently rated, the results of the studies were compared across domains as well as for the overall rating (see Table 2). The results range from moderate agreement to near perfect agreement across the domains, with domains 1, 4, 5 and 6 achieving 100% agreement. The weighted Kappa for the overall risk of bias rating on the ANROWS-IRIS is 0.795, indicating substantial agreement between assessors.

Table 2: Results from interrater reliability testing on 11 studies across two independent assessors

Domain	% Agreement	Kappa	Weighted Kappa
1. Study design	100%	1.000	1.000
2. Selection bias	82%	-0.862	0.686
3. Confounders	82%	0.645	0.676
4. Data collection methods	100%	1.000	1.000
5. Withdrawals and drop-outs	100%	1.000	1.000
6. intervention integrity and fidelity	100%	1.000	1.000
Overall rating	*73%*	*-0.407*	*0.795*

Summary

This report has detailed the approach taken to develop a bespoke risk of bias tool for application to quantitative impact evaluations that use a wide range of study designs in the social and psychological sciences, and specifically for use with the extant literature on violence against women. It described the process by which the six ANROWS-IRIS domains were established and defined the signal items for each of the domains. It also detailed how the six domains' study design, selection bias, confounders, data collection methods, withdrawals and drop-outs, and intervention integrity are combined into an overall risk of bias rating. Initial interrater reliability testing is reported, showing good agreement between assessors.

However, as with all risk of bias tools, the ANROWS-IRIS has its limitations. First, for the tool to be applied by trained non-experts and to be understood by a variety of audiences, some aspects of risk of bias were simplified in comparison to other tools. While this may reduce some nuance and specificity usually required for reviews of impact evaluation evidence, it clarified our tool to ensure accessibility. Relatedly, some items on the ANROWS-IRIS require the assessor to make subjective judgements, which raises the risk of low interrater reliability and lack of precision, depending on the assessor's expertise and depth of engagement with the study.

Second, we did not include reliability in the assessment of outcome measures. A large portion of the anticipated literature within the Evidence Portal and the field more broadly uses official or administrative data where there is not an established way of assessing reliability, unlike established self-report or questionnaire measures. An assessment of reliability in risk of bias tools is a nuanced area that does not lend itself to clear rules of thumb. Other tools rate this domain by simply asking if authors have reported reliability information; however, we do not believe that this is a true assessment of reliability and would rather be an assessment of reporting completeness.

A final issue that we grappled with during the development of the tool was how to rate studies that did not report information required for the assessment of domains – for example, formal statistical tests of differences between treatment and comparison groups and intervention integrity. To resolve this issue, we were guided by existing tools and balanced the arguments that: 1) the absence of the information cannot always be assumed to mean there are no issues; and 2) authors are often faced with barriers to reporting completeness when publishing research, such as word counts and journal specifications.

Directions for future research and implications for policy

This report showcases the initial development of the ANROWS-IRIS and its application. However, further research is warranted, both in the testing of the tool itself and in applying it to studies for the Evidence Portal and other systematic reviews.

While the initial interrater reliability testing shows good agreement between assessors and across the domains, more extensive testing is ongoing. Additionally, the project team will continue to discuss ratings at team meetings to ensure a shared understanding of the tool. Previous studies have shown that standardised training for risk of bias assessment may lead to improved reliability (e.g. da Costa et al., 2017). The project team thus plans to implement standardised and intensive training as well.

Researchers have conducted studies comparing the usability, reliability and applicability of different risk of bias tools on the systematic review "market" (e.g. Gates et al., 2018; Hartling et al., 2012; Jeyaraman et al., 2020; Kim et al., 2013). We will explore publishing a paper that contributes to this body of work by comparing the ANROWS-IRIS with similar tools for assessing risk of bias in experimental and quasi-experimental impact evaluations.

Alongside our own research, we encourage the use and continued testing of our tool by other research teams in their own systematic reviews or evidence portal projects. We also encourage those involved in designing, reporting and implementing evaluations to consider risk of bias domains as part of their protocols in order to minimise potential study flaws that may lead to over- or underestimates of the effect of the intervention. Likewise, we encourage practitioners and policymakers to consider the elements of risk of bias when funding evaluations as well as when deciding what to implement.

The ANROWS-IRIS will be invaluable to the Evidence Portal project at large. The Evidence Portal is a living resource that provides policymakers and practitioners with access to evidence on the nature and effectiveness of interventions designed to address and end violence against women in Australia and other high-income countries. As such, the Evidence Portal is strengthened by the ability to assess the risk of bias of included interventions in the field. This lends the findings from the Evidence Portal greater robustness and aims to promote evidence-informed policy- and decision-making that can account for the quality of the literature.

References

Alexander, L. K., Lopes, B., Ricchetti-Masterson, K., & Yeatts, K. B. (2015). Selection bias. In *ERIC notebook* (2nd ed., pp. 1–3). UNC Gillings School of Global Public Health. https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC13.pdf

Aupperle, R., L., Allard, C. B., Simmons, A. N., Flagan, T., Thorp, S. R., Norman, S. B., Paulus, M. P., & Stein, M. B. (2013). Neural responses during emotional processing before and after cognitive trauma therapy for battered women. *Psychiatry Research: Neuroimaging, 214*(1), 48–55. https://dx.doi.org/10.1016/j.pscychresns.2013.05.001

Beck, J. G., Tran, H. N., Dodson, T. S., Henschel, A. V., Woodward, M. J., & Eddinger, J. (2016). Cognitive trauma therapy for battered women: Replication and extension. *Psychology of Violence, 6*(3), 368–377. https://doi.org/10.1037/vio0000024

Berger, V. W. (2005). Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. *Biometrical Journal, 47*(2), 119–127. https://doi.org/10.1002/bimj.200410106

Cochrane Effective Practice and Organisation of Care. (2017). Suggested risk of bias criteria for EPOC reviews. In *EPOC Resources for review authors* (pp. 1–4). EPOC. https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/suggested_risk_of_bias_criteria_for_epoc_reviews.pdf

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

da Costa, B. R., Beckett, B., Diaz, A., Resta, N. M., Johnston, B. C., Egger, M., Jüni, P., & Armijo-Olivo, S. (2017). Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: A prospective study. *Systematic Reviews, 6*(1), 44. https://doi.org/10.1186/s13643-017-0441-7

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health, 52*(6), 377–384. https://doi.org/10.1136/jech.52.6.377

Drukker, M., Weltens, I., van Hooijdonk, C. F. M., Vandenberk, E., & Bak, M. (2021). Development of a methodological quality criteria list for observational studies: The Observational Study Quality Evaluation. *Frontiers in Research Metrics and Analytics, 6*, 675071. https://doi.org/10.3389/frma.2021.675071

Echeburúa, E., Sarasua, B., & Zubizarreta, I. (2014). Individual versus individual and group therapy regarding a cognitive-behavioral treatment for battered women in a community setting. *Journal of Interpersonal Violence, 29*(10), 1783–1801. https://doi.org/10.1177/0886260513511703

Effective Public Health Practice Project. (2009). *Quality Assessment Tool for Quantitative Studies Dictionary*. EPHPP. https://www.ephpp.ca/PDF/QADictionary_dec2009.pdf

References *continued*

Effective Public Health Practice Project. (2010). *Quality Assessment Tool for Quantitative Studies*. EPHPP. https://www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf

Eggins, E., Higginson, A., Watt, E., Hine, L., & Bell, C. (2023). *ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS): Risk of bias rating tool* (Evidence Portal tool, EP.23.04/1). ANROWS. https://evidenceportal.au/methods/#research-reports

Ferrari, G., Feder, G., Agnew-Davies, R., Bailey, J. E., Hollinghurst, S., Howard, L., Howarth, E., Sardinha, L., Sharp, D., & Peters, T. J. (2018). Psychological advocacy towards healing (PATH): A randomized controlled trial of a psychological intervention in a domestic violence service setting. *PLOS ONE, 13*(11), e0205485. https://doi.org/10.1371/journal.pone.0205485

Gates, A., Gates, M., Duarte, G., Cary, M., Becker, M., Prediger, B., Vandermeer, B., Fernandes, R. M., Pieper, D., & Hartling, L. (2018). Evaluation of the reliability, usability, and applicability of AMSTAR, AMSTAR 2, and ROBIS: Protocol for a descriptive analytic study. *Systematic Reviews, 7*(1), 85. https://doi.org/10.1186/s13643-018-0746-1

Gates, M., Gates, A., Duarte, G., Cary, M., Becker, M., Prediger, B., Vandermeer, B., Fernandes, R. M., Pieper, D., & Hartling, L. (2020). Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. *Journal of Clinical Epidemiology, 125*, 9–15. https://doi.org/10.1016/j.jclinepi.2020.04.026

Glen, S. (n.d). *Cohen's Kappa Statistic*. Retrieved August 29, 2023, from https://www.statisticshowto.com/cohens-kappa-statistic/

Hartling, L., Hamm, M., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., Tsertsvadze, A., Hempel, S., Shekelle, P., & Dryden, D. M. (2012). Validity and inter-rater reliability testing of quality assessment instruments. *Agency for Healthcare Research and Quality*. https://europepmc.org/article/MED/22536612/NBK92281#introduction.s1

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks. L., & Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ, 343*, d5928. https://doi.org/10.1136/bmj.d5928

Higgins, J. P. T., Savović, J., Page, M. J., & Sterne, J. A. C. (Eds.). (2016). *Revised Cochrane risk of bias tool for randomized trials (RoB 2.0)*. https://www.unisa.edu.au/contentassets/72bf75606a2b4abcaf7f17404af374ad/rob2-0_indiv_main_guidance.pdf

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2022). *Cochrane handbook for systematic reviews of interventions* (Version 6.3). Cochrane. https://training.cochrane.org/handbook

Higginson, A., Eggins, E., Hine, L., Watt, E., & Bell, C. (2023). *ANROWS Instrument for assessing Risk of bias in quantitative Impact Studies (ANROWS-IRIS): Risk of bias tool guidance document* (Evidence Portal guidance document, EP.23.03/1). ANROWS. https://evidenceportal.au/methods/#research-reports

JBI. (2020). *Checklist for quasi-experimental studies (non-randomized experimental studies): Critical appraisal tools for use in JBI systematic reviews*. JBI. https://jbi.global/sites/default/files/2020-07/Checklist_for_Quasi-Experimental_Appraisal_Tool.pdf

Jeyaraman, M. M., Rabbani, R., Copstein, L., Robson, R. C., Al-Yousif, N., Pollock, M., Xia, J., Balijepalli, C., Hofer, K., Mansour, S., Fazeli, M. S., Ansari, M. T., Tricco, A. C., & Abou-Setta, A. M. (2020). Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. *Journal of Clinical Epidemiology, 128*, 140–147. https://doi.org/10.1016/j.jclinepi.2020.09.033

References *continued*

Johnson, D. M., Johnson, N. L., Perez, S. K., Palmieri, P. A., & Zlotnick, C. (2016). Comparison of adding treatment of PTSD during and after shelter stay to standard care in residents of battered women's shelters: Results of a randomized clinical trial. *Journal of Traumatic Stress, 29*(4), 365–373. https://doi.org/10.1002/jts.22117

Johnson, D. M., Zlotnick, C., & Perez, S. (2011). Cognitive behavioral treatment of PTSD in residents of battered women's shelters: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 79*(4), 542–551. https://doi.org/10.1037/a0023822

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA: Journal of the American Medical Association, 282*(11), 1054–1060. https://doi.org/10.1001/jama.282.11.1054

Kim, S. Y., Park, J. E., Lee, Y. J., Seo, H.-J., Sheen, S.-S., Hahn, S., Jang, B.-H., & Son, H.-J. (2013). Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology, 66*(4), 408–414. https://doi.org/10.1016/j.jclinepi.2012.09.016

Littleton, H., & Grills, A. (2019). Changes in coping and negative cognitions as mechanisms of change in online treatment for rape-related posttraumatic stress disorder. *Journal of Traumatic Stress, 32*(6), 927–935. https://doi.org/10.1002/jts.22447

Littleton, H., Grills, A. E., Kline, K. D., Schoemann, A. M., & Dodd, J. C. (2016). The From Survivor to Thriver program: RCT of an online therapist-facilitated program for rape-related PTSD. *Journal of Anxiety Disorders, 43*, 41–51. https://doi.org/10.1016/j.janxdis.2016.07.010

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Moreira, A., Moreira, A. C., & Rocha, J. C. (2022). Randomized controlled trial: Cognitive-narrative therapy for IPV victims. *Journal of Interpersonal Violence*, *37*(5–6), NP2998–NP3014. https://doi.org/10.1177/0886260520943719

National Institute for Health and Clinical Excellence. (2012). *Methods for the development of NICE public health guidance: Process and methods* (3rd ed.). NICE. https://www.nice.org.uk/process/pmg4

Page, M. J., McKenzie, J. E., & Higgins, J. P. T. (2018). Tools for assessing risk of reporting biases in studies and syntheses of studies: A systematic review. *BMJ Open, 8*(3), e019703. https://doi.org/10.1136/bmjopen-2017-019703

Quigley, J. M., Thompson, J. C., Halfpenny, N. J., & Scott, D. A. (2019). Critical appraisal of nonrandomized studies – a review of recommended and commonly used tools. *Journal of Evaluation in Clinical Practice, 25*(1), 44–52. https://doi.org/10.1111/jep.12889

Scottish Intercollegiate Guidelines Network, & Healthcare Improvement Scotland. (2011). *Sign 50: A guideline developer's handbook* (Revised ed.). SIGN. https://www.sign.ac.uk/assets/sign50_2011.pdf

Seehra, J., Pandis, N., Koletsi, D., & Fleming, P. S. (2016). Use of quality assessment tools in systematic reviews was varied and inconsistent. *Journal of Clinical Epidemiology, 69*, 179–184. https://doi.org/10.1016/j.jclinepi.2015.06.023

References *continued*

Smith, E. N., Rotunda, R. J, & Cosio-Lima, L. (2015). Cognitive behavioral therapy and aerobic exercise for survivors of sexual violence with posttraumatic stress disorder: A feasibility study. *Journal of Traumatic Stress Disorders & Treatment, 4*(1), 1–6. https://doi.org/10.4172/2324-8947.1000136

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott . . . Higgins, J. P. T. (2016a). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Higgins, J. P. T., Elbers, R. G., Reeves, B. C., & the development group for ROBINS-I. (2016b). *Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): Detailed guidance*. https://www.riskofbias.info/welcome/home/current-version-of-robins-i/robins-i-detailed-guidance-2016

The Campbell Collaboration. (2021). *Campbell systematic reviews: Policies and guidelines* (Version 1.8). The Campbell Collaboration. https://doi.org/10.4073/cpg.2016.1

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176–184. https://doi.org/10.1111/j.1524-475X.2004.04006.x

Waddington, H., Aloe, A. M., Becker, B. J., Djimeu, E. W., Hombrados, J. G., Tugwell, P., Wells, G., & Reeves, B. (2017). Quasi-experimental study designs series – paper 6: Risk of bias assessment. *Journal of Clinical Epidemiology, 89*, 43–52. https://doi.org/10.1016/j.jclinepi.2017.02.015

Wells, S. Y., Glassman, L. H., Talkovsky, A. M., Chatfield, M. A., Sohn, M. J., Morland, L. A., & Mackintosh, M.-A. (2019). Examining changes in sexual functioning after cognitive processing therapy in a sample of women trauma survivors. *Women's Health Issues, 29*(1), 72–79. https://doi.org/10.1016/j.whi.2018.10.003

White, H., Saran, A., Fowler, B., Portes, A., Fitzpatrick, S., & Teixeira, L. (2020). PROTOCOL: Studies of the effectiveness of interventions to improve the welfare of those affected by, and at risk of, homelessness in high-income countries: An evidence and gap map. *Campbell Systematic Reviews, 16*(1), e1069. https://doi.org/10.1002/cl2.1069

ANROWS

AUSTRALIA'S NATIONAL RESEARCH
ORGANISATION FOR WOMEN'S SAFETY
to Reduce Violence against Women & their Children