National Institute for Health and Care Excellence

Draft for consultation

Mental wellbeing at work

NICE guideline: methods

NICE guideline NGXX

Methods

September 2021

Draft for Consultation

Evidence reviews were developed by Public Health – Internal Guideline Development team

Disclaimer

The recommendations in this guideline represent the view of NICE, arrived at after careful consideration of the evidence available. When exercising their judgement, professionals are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or service users. The recommendations in this guideline are not mandatory and the guideline does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and/or their carer or guardian.

Local commissioners and/or providers have a responsibility to enable the guideline to be applied when individual health professionals and their patients or service users wish to use it. They should do so in the context of local and national priorities for funding and developing services, and in light of their duties to have due regard to the need to eliminate unlawful discrimination, to advance equality of opportunity and to reduce health inequalities. Nothing in this guideline should be interpreted in a way that would be inconsistent with compliance with those duties.

NICE guidelines cover health and care in England. Decisions on how they apply in other UK countries are made by ministers in the Welsh Government, Scottish Government, and Northern Ireland Executive. All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN:

Contents

Development of the guideline

Remit

To see "What this guideline covers" and "What this guideline does not cover" please see the guideline scope Mental wellbeing at work.

Methods

This guideline was developed using the methods described in the NICE guidelines manual as outlined in the table below. Scoping was carried out using the 2014 version of the NICE manual. The remainder of the development followed the 2020 version of the NICE manual.

Where the guidelines manual does not provide advice, additional methods are described below.

Developing the review questions and outcomes

The 14 review questions developed for this guideline are addressed in six evidence reviews, and were based on the key areas identified in the guideline scope. Review questions were developed by the NICE Public Health Internal Guideline Development (PHIGD) team and refined, validated and signed off by the Public Health Advisory Committee (PHAC) and NICE quality assurance team.

The qualitative and quantitative review questions were based on the population, intervention, comparator and outcome (PICO) framework.

Full literature searches, critical appraisals and evidence reviews were completed for all review questions. Details of these elements are found in the review protocols for each review (see Appendix A of each relevant review). Where protocol deviations have been made, these will be reported in the Methods section of the individual review.

Table 1: Summary of review questions and index to evidence reviews

Evidence review	Review questions	Type of review
A	1.1 What universal, organisational-level interventions, programmes, policies or strategies are effective and cost effective at: • promoting positive mental wellbeing? • improving mental wellbeing? • preventing poor mental wellbeing? 1.2 What interventions or strategies effectively and cost-effectively help employers and peers • to recognise and engage employees who may require support for their mental wellbeing, or • to identify periods of high risk within an organisation?	Mixed methods review

Evidence review	Review questions	Type of review
	1.3 For the following groups in relation to organisational-level targeted interventions, what are their views and experiences of what and why certain approaches may or may not work, and how it could be improved: • employees receiving them • employers • those delivering them?	
B	2.1 What training to help managers to understand, promote and support mental wellbeing is effective and cost-effective? 2.2 What training is effective and cost-effective to help managers to improve their knowledge and skills in recognising employees who experience or are at risk of poor mental wellbeing? 2.3 What training is effective and cost-effective in helping managers to improve their knowledge and skills in responding to mental wellbeing issues? 2.4 For the following groups in relation to approaches to training managers in employee mental wellbeing, what are their views and experiences of what and why certain approaches may or may not work, and how it could be improved: • managers receiving them • employees who will interact with managers • employers • those delivering them?	Mixed methods review
C	3.1 What, organisational-level interventions, programmes, policies or strategies targeted to employees who experience or are identified as being at risk of poor mental wellbeing at work are effective and cost effective at: • promoting positive mental wellbeing? • improving mental wellbeing? • preventing poor mental wellbeing? 3.2 For the following groups in relation to organisational-level targeted interventions, what are their views and experiences of what and why certain approaches may or may not work, and how it could be improved: • employees receiving them • employers • those delivering them?	Mixed methods review
D	4.1 What universal, individual-level interventions, programmes, policies or strategies are effective and cost effective at:	–Mixed methods review

Evidence review	Review questions	Type of review
	• promoting positive mental wellbeing? • improving mental wellbeing? • preventing poor mental wellbeing? 4.2 For the following groups in relation to universal individual-level interventions, what are their views and experiences of what and why certain approaches may or may not work, and how it could be improved: • those receiving them • employers • those delivering them?	
E	5.1 What individual-level interventions targeted to employees who experience, or are identified as being at risk of, poor mental wellbeing at work are effective and cost effective for: • promoting positive mental wellbeing? • improving mental wellbeing? • preventing poor mental wellbeing? 5.2 For the following groups in relation to individual-level targeted interventions, what are their views and experiences of what and why certain approaches may or may not work, and how it could be improved: • those receiving them? • employers? • those delivering them?	–Mixed method review
F	6.1 What are the barriers and facilitators to, and key aspects of (including systems and processes), the successful implementation or delivery of mental wellbeing interventions, programmes, policies or strategies at work?	Qualitative

The COMET database was searched for core outcome sets relevant to this guideline. no core outcome sets were identified at the time of this search and therefore the outcomes for evidence reviews were based on committee discussions.

Reviewing research evidence

Review protocols

Review protocols were developed with the guideline committee to outline the inclusion and exclusion criteria used to select studies for each evidence review. Where possible, review protocols were prospectively registered in the PROSPERO register of systematic reviews. Five of the six review protocols were registered in the PROSPERO register of systematic reviews. Review protocols are published in Appendix A of each review with the PROSPERO registration number where available.

Protocol deviations

Reporting of continuous outcomes

The committee discussed the presentation of continuous outcomes from the included studies in all of the evidence reviews. They noted that studies of different interventions had used different scales to measure the same outcome making it difficult to informally compare the size of effect across different interventions. They noted that it was also confusing that some meta-analyses (using the same outcome measure) were reported using mean difference and those where the outcomes were mixed used standardised mean difference (SMD).

The committee considered possible solutions for this and agreed that it would be informative to standardise all of mean differences reported in the studies for all outcomes. The committee were satisfied the key assumption underlying SMDs - that the differences in standard deviations among studies reflect differences in measurement scales and not real differences in variability among study populations – was met. The committee also agreed this approach would have an additional benefit in the economic model ensuring that the inputs relating to effectiveness were standardised should users want to compare the impact of interventions using the same outcome.

The committee were aware of the potential inaccuracies that can arise from this, and of the difficulties in interpreting standardised mean differences, however they agreed that the line of no effect was still a suitable cut off for determining whether there was a meaningful effect, and that magnitude of effect could be judged using the rules of thumb described by Cohen that suggest that an SMD of 0.2 represents a “small” effect, an SMD of 0.5 represents a “medium” effect, and an SMD of 0.8 represents a “large” effect[1]. This method is also judged to be a useful way to understand the magnitude of effect when unfamiliar scales are being used[2]. For example, when you read that a treatment group’s mean post-treatment score on scale X was 10 points higher than that of a control group, there is no way of appreciating how much a difference this actually represents unless you are very familiar with the scale that is being used. But if the difference is expressed in terms of SMD as corresponding to an effect size of 0.5, for example, you can understand that it represents a moderate effectiveness in comparison with the control. In fact, Tian et al[3]. noted that the SMD does not depend on the unit of measurement, and therefore the SMD has been widely used as a measure of intervention effect in many applied field[4]s.

A study by Takeshima et al concluded that The SMD was more generalizable than the MD. The MD had a greater statistical power than the SMD but did not result in material differences[4].

1 Cohen J: Statistical Power Analysis for the Behavioral Sciences. 1988, Hillsdale, New Jersey: Lawrence Erlbaum Associates: Routledge

2 Borenstein M, Hedges LV, Higgins JP, Rothstein HR: Introduction to Meta-Analysis. 2011, Wiley.com

3 Tian L: Inferences on standardized mean difference: the generalized variable approach. Stat Med. 2007, 26 (5): 945-953. 10.1002/sim.2589

4 Takeshima, N., Sozu, T., Tajika, A. et al. Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference?. BMC Med Res Methodol 14, 30 (2014). https://doi.org/10.1186/1471-2288-14-30

Priority screening

For the combined search for reviews A, C, D, E the figure of 60%of records screened was not achieved, and a pragmatic decision was made to stop screening at approximately 25%. This decision was based on the size of the search (72, 259 records) and the screening of over 7000 titles and abstracts with no additional includes.

Searching for evidence

Evidence was searched for each review question using the methods specified in the 2018 NICE guidelines manual. Full details of search strategies, databases searched and numbers of studies identified can be found in the appendices of each individual review.

Selecting studies for inclusion

All references identified by the literature searches and from other sources (for example, previous versions of the guideline or studies identified by committee members) were uploaded into EPPI reviewer software (version 5) and de-duplicated. Titles and abstracts were assessed for possible inclusion using the criteria specified in the review protocol. 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

The evidence reviews made use of the priority screening functionality within the EPPI-reviewer software. This functionality uses a machine learning algorithm (specifically, an SGD classifier) to take information on features (1-, 2- and 3-word blocks) in the titles and abstract of papers marked as being ‘includes’ or ‘excludes’ during the title and abstract screening process, and re-orders the remaining records from most likely to least likely to be an include, based on that algorithm. This re-ordering of the remaining records occurs every time 25 additional records have been screened. Research is currently ongoing as to what are the appropriate thresholds where reviewing of abstracts can be stopped, assuming a defined threshold for the proportion of relevant papers it is acceptable to miss on primary screening. As a conservative approach until that research has been completed, the protocol for reviews in this guideline specified screening at least 60% of the identified abstracts (or 1,000 records, if that is a greater number).

As an additional check to ensure this approach did not miss relevant studies, systematic reviews were included in the review protocol and search strategy for all review questions. Relevant systematic reviews or qualitative evidence syntheses were used to identify any papers not found through the primary search. Committee members were also consulted to identify studies that were missed. The protocol outlines that if additional studies were found that were erroneously excluded during the priority screening process, the full database would be subsequently screened, however this did not occur.

The full text of potentially eligible studies was retrieved and assessed according to the criteria specified in the review protocol and screened against the protocol at full text to determine final included studies. A standardised form was used to extract data from included studies into the EPPI-R5 reviewer software.

Methods of combining evidence

Data synthesis for intervention studies

Where possible, meta-analyses were conducted to combine the results of quantitative studies for each outcome.

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane Review Manager V5.3 where possible. Meta-analyses that could not be conducted in Cochrane Review Manager were carried out in R version 3.3.4. using the package 'metafor'. A pooled relative risk was calculated for dichotomous outcomes (using the Mantel–Haenszel method) reporting numbers of people having an event. Both relative and absolute risks were presented, with absolute risks calculated by applying the relative risk to the risk in the comparator arm of the meta-analysis (calculated as the total number of events in the comparator arms of studies in the meta-analysis divided by the total number of participants in the comparator arms of studies in the meta-analysis).

Continuous outcomes were all standardised to the same scale before meta-analysis was conducted on the mean differences (see protocol deviation above).

For continuous outcomes analysed as standardised mean differences, change from baseline values were used in the meta-analysis if they were accompanied by a measure of spread (for example standard deviation). Where change from baseline (accompanied by a measure of spread) were not reported, the corresponding values at the timepoint of interest were used. If only a subset of trials reported change from baseline data, final timepoint values were combined with change from baseline values to produce summary estimates of effect.

Random effects models were fitted when there was significant between-study heterogeneity in methodology, population, intervention or comparator was identified by the reviewer in advance of data analysis. This decision was made and recorded before any data analysis was undertaken.

Cluster randomised controlled trials

Where cluster randomised controlled trials have been pooled with individually randomised controlled trials, the number of people included in the analysis from these trials have been adjusted using a reported or imputed intra-class correlation coefficient (ICC) for that outcome. When the studies did not report the ICC and we could not impute it, we included the study data without adjustment and noted this in the evidence table.

In some studies, the unit of randomization was the individual and in some the unit was a cluster (workplace). When the unit of randomization was the cluster then outcome data at an individual level was adjusted for cluster effect as outlined above and in the Cochrane Handbook for Systematic Reviews of Interventions.

Data synthesis for qualitative reviews

Where multiple qualitative studies were identified for a single question, information from the studies was combined using a thematic synthesis. The thematic synthesis was based partly on a priori categories describing phenomena the committee was

interested in (for example, using an existing model [framework synthesis]) and partly on themes that emerged from the coding of the included studies. Papers were uploaded to MS Excel where the relevant data from the papers were coded. Once all of the included studies had been examined and coded, the resulting sets of codes were aggregated into themes and sub-themes and were evaluated using CERQual. The aggregated themes were used to develop interpretive 'review findings'. These review findings were reproduced in a summary of qualitative findings table along with example quotes and details of the CERQual assessment of each review finding.

Data synthesis for mixed methods reviews

Data synthesis for mixed methods reviews was carried out in accordance with the Joanna Briggs Institute manual for evidence synthesis (https://wiki.jbi.global/display/MANUAL) chapter 8. Synthesis followed a convergent segregated approach where independent synthesis of quantitative data and qualitative data was undertaken, followed by the integration of the two types of evidence.

The qualitative and quantitative reviews were presented separately in the reviews and an integration section was written that addressed the following questions:

- Are the results/findings from individual syntheses supportive or contradictory?
- Does the qualitative evidence explain why the intervention is/is not effective?
- Does the qualitative evidence explain differences in the direction and size of effect across the included quantitative studies?
- Which aspects of the quantitative evidence were/were not explored in the qualitative studies?
- Which aspects of the qualitative evidence were/were not tested in the quantitative studies?

Where appropriate, and data from quantitative and qualitative sections of the review were integrated into tables or logic models/conceptual frameworks to show possible interrelationships between them.

Appraising the quality of evidence

Intervention studies (relative effect estimates)

RCTs and quasi-randomised controlled trials were quality assessed using the Cochrane Risk of Bias Tool. Non-randomised controlled trials and cohort studies were quality assessed using the ROBINS-I tool. Other study types (for example controlled before and after studies) were assessed using the preferred option specified in the NICE guidelines manual 2018 (appendix H). Evidence on each outcome for each individual study was classified into one of the following groups:

- **Low risk of bias** – The true effect size for the study is likely to be close to the estimated effect size.
- **Moderate risk of bias** – There is a possibility the true effect size for the study is substantially different to the estimated effect size.

- **High risk of bias** – It is likely the true effect size for the study is substantially different to the estimated effect size.
- **Critical risk of bias** (ROBINS-I only) - It is very likely the true effect size for the study is substantially different to the estimated effect size.

Minimally important differences (MIDs) and decision thresholds

The Core Outcome Measures in Effectiveness Trials (COMET) database was searched to identify published minimal important difference thresholds relevant to this guideline that might aid the committee in identifying decision thresholds for the purpose of GRADE. Identified MIDs were assessed to ensure they had been developed and validated in a methodologically rigorous way, and were applicable to the populations, interventions and outcomes specified in this guideline. In addition, PHAC members were asked to prospectively specify any outcomes where they felt a consensus decision threshold could be defined from their experience.

Decision thresholds were used to assess imprecision using GRADE and aid interpretation of the size of effects for different outcomes. No published decision thresholds were found so the line of no effect was used for the purposes of identifying meaningful effect and imprecision.

Public health decision thresholds

The committee were asked to define decision thresholds for association outcomes based on the degree of association that was considered important for decision making. The committee were unable to define a decision threshold by consensus for outcomes of interest. The decision was made not to use statistically calculated MIDs of 0.8 and 1.25 for relative risk and 50% of the median SD of the control groups for MDs because these are clinical thresholds and do not have meaning in public health interventions where any effect is considered to be potentially significant so the line of no effect was used at the decision threshold for the purpose of rating imprecision in GRADE.

GRADE for pairwise meta-analyses of interventional evidence

GRADE was used to assess the quality of evidence for the outcomes specified in the review protocol. Data from randomised controlled trials, non-randomised controlled trials and cohort studies (which were quality assessed using the Cochrane risk of bias tool or ROBINS-I) were initially rated as high quality while data from other study types were initially rated as low quality. The quality of the evidence for each outcome was downgraded or not from this initial point, based on the criteria given in Table 2.

Table 2: Rationale for downgrading quality of evidence for intervention studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate high or critical risk of bias, the outcome was downgraded one level.

GRADE criteria	Reasons for downgrading quality
	Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high or critical risk of bias, the outcome was downgraded two levels.
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study. Not serious: If the I^2 was less than 50%, the outcome was not downgraded. Serious: If the I^2 was greater than 50%, the outcome was downgraded one level, if the I^2 was greater than 75%, the outcome was downgraded two levels
Imprecision	The line of no effect was defined as an MID for the outcome, it was downgraded once if the 95% confidence interval for the effect size crossed the line of no effect (i.e. the outcome was not statistically significant).
Publication bias	Publication bias was not assessed because no meta-analyses involved more than 10 studies which is considered the minimum for a meaningful funnel plot..

Qualitative studies

Individual qualitative studies were critically appraised using the CASP qualitative checklist. Each individual study was classified into one of the following three groups:

1. **No/minor concerns** – The findings and themes identified in the study are likely to accurately capture the true picture.
2. **Moderate concerns** – There is a possibility the findings and themes identified in the study are not a complete representation of the true picture.
3. **Serious concerns** – It is likely the findings and themes identified in the study are not a complete representation of the true picture

CERQual was used to assess the confidence we have in each of the review findings. Evidence from all qualitative study designs (interviews, focus groups etc.) was initially rated as high confidence and the confidence in the evidence for each theme was then downgraded from this initial point as detailed in Table 3 below.

CERQual has been applied at theme level in reviews A to F where all of the sub-themes associated with a theme occur in the same papers (or where there are no sub-themes). The exception is review F, where due to the nature of the review, sub-themes are reported in different papers than other sub-themes in the theme. As a

result, CERQual was applied at sub-theme level in this review because it would not necessarily be consistent across the sub-themes.

Table 3 Rationale for downgrading confidence in evidence for qualitative questions

CERQual criteria	Reasons for downgrading confidence
Methodological limitations	No/Low concerns: If the theme was identified in studies at low risk of bias, the outcome was not downgraded Moderate concerns: If the theme was identified only in studies at moderate or high risk of bias, the outcome was downgraded one level. Serious concerns: If the theme was identified only in studies at high risk of bias, the outcome was downgraded two levels.
Relevance	High: If the theme was identified in highly relevant studies only or highly relevant and relevant studies only, the outcome was not downgraded Moderate: If the theme was identified only in relevant and partially relevant studies, the outcome was downgraded one level. Low: If the theme was identified only in partially relevant studies, the outcome was downgraded two levels.
Coherence	Coherence was addressed based on two factors: Between study – does the theme consistently emerge from all relevant studies Theoretical – does the theme provide a convincing theoretical explanation for the patterns found in the data The outcome was downgraded once if there were concerns about one of these elements of coherence, and twice if there were concerns about both elements.
Adequacy of data	The outcome was downgraded if there was insufficient data to develop an understanding of the phenomenon of interest, either due to insufficient studies, participants, observations, or the complexity of the theme.
Overall confidence rating	The confidence or certainty in the theme is classified as high, moderate, low or very low. Our assessment of confidence communicates the extent to which the research finding is likely to be substantially different from the phenomenon of interest. By substantially different, we mean different enough that it might change how the finding influences a practical or policy decision about health, social care, or other interventions. High: It is highly likely that the review finding is a reasonable representation of the phenomenon of interest Moderate: It is likely that the review finding is a reasonable representation of the phenomenon of interest Low: It is possible that the review finding is a reasonable representation of the phenomenon of interest Very Low: It is unclear whether the review finding is a reasonable representation of the phenomenon of interest

Mixed methods studies

Mixed methods studies were evaluated using the appropriate quality assessment tools for the component study types, see sections on intervention studies and qualitative studies. Other methods of assessing mixed methods studies were agreed

with the NICE methods and economics team QA lead and reported in the individual reviews.

Reviewing economic evidence

Inclusion and exclusion of economic studies

Literature reviews seeking to identify published cost–utility analyses of relevance to the issues under consideration were conducted for all questions. In each case, the search undertaken for the public health review was modified, retaining population and intervention descriptors, but removing any study-design filter and adding a filter designed to identify relevant health economic analyses. In assessing studies for inclusion, population, intervention and comparator, criteria were always identical to those used in the parallel public health search; only cost–utility analyses were included. Economic evidence profiles, including critical appraisal according to the Guidelines manual, were completed for included studies.

Appraising the quality of economic evidence

Economic studies identified through a systematic search of the literature were appraised using a methodology checklist designed for economic evaluations (NICE guidelines manual; 2018, 2020). This checklist is not intended to judge the quality of a study per se, but to determine whether an existing economic evaluation is useful to inform the decision-making of the committee for a specific topic within the guideline.

There are 2 parts of the appraisal process. The first step is to assess applicability (that is, the relevance of the study to the specific guideline topic and the NICE reference case); evaluations are categorised according to the criteria in Table 4.

Table 4 Applicability criteria

Level	Explanation
Directly applicable	The study meets all applicability criteria, or fails to meet one or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness
Partially applicable	The study fails to meet one or more applicability criteria, and this could change the conclusions about cost effectiveness
Not applicable	The study fails to meet one or more applicability criteria, and this is likely to change the conclusions about cost effectiveness. These studies are excluded from further consideration

In the second step, only those studies deemed directly or partially applicable are further assessed for limitations (that is, methodological quality); see categorisation criteria in Table 5.

Table 5 Methodological criteria

Level	Explanation
Minor limitations	Meets all quality criteria, or fails to meet one or more quality criteria but this is unlikely to change the conclusions about cost effectiveness

Level	Explanation
Potentially serious limitations	Fails to meet one or more quality criteria and this could change the conclusions about cost effectiveness
Very serious limitations	Fails to meet one or more quality criteria and this is highly likely to change the conclusions about cost effectiveness. Such studies should usually be excluded from further consideration

Where relevant, a summary of the main findings from the systematic search, review and appraisal of economic evidence is presented in an economic evidence profile alongside the public health evidence.

Health economic modelling

As well as reviewing the published economic literature for each review question, as described above, de novo economic analysis was undertaken in selected areas. Priority areas for new health economic analysis were agreed by the committee.

The following general principles were adhered to in developing the analysis:

- Methods were consistent with the NICE reference case.
- The design of the model, selection of inputs and interpretation of the results was discussed and agreed with the committee.
- Where possible, model inputs were based on the systematic review of the public health literature, supplemented with other published data sources identified by the committee as required.
- When published data were not available committee expert opinion was used to populate the model.
- Model inputs and assumptions were reported fully and transparently.
- The results were subject to sensitivity analysis and limitations were discussed.

Full methods for the de novo cost-effectiveness analysis are described in the HE report.

Resource impact assessment

The resource impact team used the methods outlined in the in Assessing resource impact process manual: guidelines

The resource impact team worked with the guideline committee from an early stage to identify recommendations that either individually or cumulatively would have a substantial impact on resources. The aim was to ensure that a recommendation would not introduce a cost pressure into the health and social care system unless the committee was convinced of the benefits and cost effectiveness of the recommendation. The team gave advice to the committee on issues related to the workforce, capacity and demand, training, facilities and educational implications of the recommendations.