Organizing Trustworthiness Assessments

HS24

This assessment organisation tool was adapted from a draft of Wilkinson et al.’s ([2024](https://bmjopen.bmj.com/content/14/3/e084164)) INSPECT-SR (Inspecting systematic reviews in health research for trustworthiness), specifically the July 2024 draft. Whereas INSPECT-SR is focused on trustworthiness assessments for health RCTs, this is adapted with psychology research in mind.

Text highlighted in yellow are Jack Wilkinson’s early notes on guidance for the checks that require more expansion. Text highlighted in blue and the ‘proposals’ section onwards are additions by Ian.

# Summary of checks

| Domain | Check | Response |
| --- | --- | --- |
| Inspecting post-publication notices | [1.1 Has the study been retracted or does it have an expression of concern, or a relevant post-publication amendment?](#_gjdgxs) | * Yes * No * Unclear * NA |
| [1.2 Do other studies by the research team highlight causes for concern (critical retraction, expression of concern, relevant post-publication amendments?)](#_30j0zll) | * Yes * No * Unclear * NA |
| Inspecting conduct, governance and transparency | [2.1 Are there concerns relating to the timing or absence of study registration?](#_1fob9te) | * Yes * No * Unclear * NA |
| [2.2 Are there important inconsistencies between the publication and the registration documents?](#_3znysh7) | * Yes * No * Unclear * NA |
| [2.3 Are there concerns relating to ethical approval?](#_2et92p0) | * Yes * No * Unclear * NA |
| [2.4 Is the recruitment and follow-up of participants implausible?](#_tyjcwt) | * Yes * No * Unclear * NA |
| [2.5 Are there concerns about the plausibility of conducting the study using the reported methods and resources?](#_3dy6vkm) | * Yes * No * Unclear * NA |
| Inspecting text and figures | [3.1 Are there concerns relating to text that is incompatible with the study?](#_1t3h5sf) | * Yes * No * Unclear * NA |
| [3.2 Is there evidence of manipulation or duplication of images?](#_4d34og8) | * Yes * No * Unclear * NA |
| Inspecting results in the paper | [4.1 Are there any discrepancies between reported data and participant inclusion criteria?](#_2s8eyo1) | * Yes * No * Unclear * NA |
| [4.2 Are any baseline data implausible?](#_17dp8vu) | * Yes * No * Unclear * NA |
| [4.3 Are there any discrepancies between results reported in figures, tables and text?](#_3rdcrjn) | * Yes * No * Unclear * NA |
| [4.4 Are there concerns relating to the plausibility of numbers of participants lost to follow-up?](#_26in1rg) | * Yes * No * Unclear * NA |
| [4.5 Are numbers of participants inconsistent throughout the publication?](#_lnxbz9) | * Yes * No * Unclear * NA |
| [4.6 Are any outcome data, including estimated treatment effects, implausible?](#_35nkun2) | * Yes * No * Unclear * NA |
| [4.7 Are the means and variances of integer data impossible?](#_1ksv4uv)  e.g. GRIM/MER + TIDES | * Yes * No * Unclear * NA |
| [4.8 Are there errors or inconsistencies in statistical analyses?](#_44sinio)  e.g. StatCheck | * Yes * No * Unclear * NA |
| [4.9 Are any other contradictions implied by the data?](#_2jxsxqh)  e.g. PORT and ANCHOR | * Yes * No * Unclear * NA |
| [4.10 Are there inconsistencies in descriptions of methods and results across publications describing the study?](#_z337ya) | * Yes * No * Unclear * NA |
| Overall study judgement \* |  | * No concerns * Some concerns * Serious concerns |
| \* *Notes:* currently there is not a quantitative scoring system for integrating the above items into this overall judgement. For the moment, this should be done subjectively. | | |

# Additional guidance for checks

These points will eventually be fleshed out into detailed guidance – current development of INSPECT-SR is focused on the selection of the questions rather than guidance on how to answer them.

## Inspecting post-publication notices

### 

### Has the study been retracted or does it have an expression of concern, or a relevant post-publication amendment?

* Check journal website and Retraction Watch (<https://retractionwatch.com/>).
* If the article is listed as retracted on the journal’s website for that article, stop assessment and label the article as untrustworthy.
* If there is an Expression of Concern (EoC) listed on the journal’s website for the article, suggesting study under investigation, put in “awaiting assessment”, which means it should also be excluded as potentially untrustworthy.
* Reviewers may find it helpful to consult sources such as PubPeer (<https://pubpeer.com>), as feedback may be visible much sooner. Expressions of concern and retractions take a long time. We suggest that criticism on PubPeer might be best used to assist in the reviewer’s assessment of the study using INSPECT-SR (for example, by directing attention to a problematic feature which can be incorporated into the corresponding domain assessment).
* Is the time between submission and publication implausibly short? See Dorothy Bishop’s [examples](https://deevybee.blogspot.com/2016/06/editorial-integrity-publishers-on-front.html) of some editors and journals routinely accepting articles within days, implying no peer review was conducted.

### 

### Do other studies by the research team highlight causes for concern (critical retraction, expression of concern, relevant post-publication amendments?)

* Search Retraction Watch (<https://retractionwatch.com/>).
* Suggest search first and last author (at minimum? Or make point that previous misconduct by a middle author might be less relevant?)
* A track record of problems might introduce doubts about the index study. Could we have serious concerns (no need to check further) on the basis of this one alone? Probably not – would mean that the study was never looked at.
* How to handle PubPeer comments on other studies? Use to direct attention to potentially problematic features in the current study?

## Inspecting conduct, governance and transparency

### 

### Are there concerns relating to the timing or absence of study registration?

* Need to come up with a definition of what counts as concerning re: timing of registration. E.g. registering one month after start of recruitment for a long study probably okay.
* Do we have a cut-off re: years, e.g. 2010? No expectation to register before a certain time.
* Could problematic reg alone lead to serious concerns? E.g. would we want to say that a trial without registration does not need to be assessed further, but could be excluded (as is the case for RIA, for example?) Would have a major impact given half of trials not registered.

### 

### Are there important inconsistencies between the publication and the registration documents?

* Need to check history of changes (details may have been changed to match the paper).
* We do not include outcome switching, which would be covered by bias assessments.

### 

### Are there concerns relating to ethical approval?

* Was ethical approval received, **and** was the ethics committee certified? [How to ascertain?]

### 

### Is the recruitment and follow-up of participants implausible?

* Plausibility of recruitment in the reported timeframe – requires domain knowledge.
* ‘Follow-up’ is going to make people look at drop out rates here (covered elsewhere). Remove reference to ‘follow-up’ and make sure this is covered in guidance for next check?

### 

### Are there concerns about the plausibility of conducting the study using the reported methods and resources?

* Plausibility of methods/ protocol. Consider reported funding/ staffing.

## Inspecting text and figures

### 

### Are there concerns relating to text that is incompatible with the study?

* Is there any text that does not correspond to the study? [need a good example here]

### 

### Is there evidence of manipulation or duplication of images?

* Visual check – any signs of tampering with plots of data
* Not recommending use of any particular software here

## Inspecting results in the paper

### 

### Are there any [unexplained] discrepancies between reported data and participant inclusion criteria?

* Need to be clear that this refers to unexplained discrepancies – perhaps this should be added to the name of the check.

### 

### Are any baseline data implausible?

* Magnitude, frequency, variance, repeated values, recurrence of even/ odd numbers, zeroes, fives etc
* In some cases, it might be useful to consider the distribution of p-values corresponding to balance checks. Limitations include the fact that many studies will not present enough characteristics to make a clear assessment. Consideration of the baseline p-values cannot be reduced to a comparison against a uniform distribution, without consideration of other explanations for non-uniform p-values (including rounding of summary data, test assumptions, use of stratified randomisation or minimisation). If these analyses are performed, it is preferable to use p-values provided by authors, rather than recalculating these based on summary data.

### 

### Are there any discrepancies between results reported in figures, tables and text?

* Check for contradictions where the same results are reported in multiple places (figures, tables, main text, abstract).
  + Between plots that repeat information, eg forest plot and funnel plot in a meta-analysis.
* Option to include protocol/SAP/protocol papers here?

### 

### Are there concerns relating to the plausibility of numbers of participants lost to follow-up?

* May require domain knowledge. Are low numbers of loss to follow-up plausible?

### 

### Are numbers of participants inconsistent throughout the publication?

* Are there unexplained inconsistencies in numbers of participants reported in different parts of the manuscript?
* Need to take care not to mistake differences in numbers due to e.g. loss to follow-up, exclusion of participants due to non-adherence

### 

### Are any outcome data, including estimated treatment effects, implausible?

* Similar considerations to baseline check – magnitude, frequency, variance of outcome values
  + In addition, plausibility of effect estimate – caution against using point estimate without reference to e.g. confidence intervals.
  + Comparison by plotting against other studies in a forest plot may be useful.
* Many metrics can be checked: standardized effect size (e.g., Cohen’s d, Pearson’s *r* correlation), unstandardized effect size (note that it may be possible to convert between standardized and unstandardized to get more info), reliability, *p* values, or others.
  + See Ian’s tool “Intuitions for Magnitudes”: <https://errors.shinyapps.io/intuitions_for_magnitudes/> (currently semi-broken)
  + Or his simpler tool displaying just Cohen’s d effect sizes: <https://errors.shinyapps.io/effect_size_percentiles/>
    - Note that both of these are merely illustrative: Effect sizes in one domain (e.g., depression interventions) may not be representative of those in other domains.
  + Sometimes when abbreviated p values (e.g., “< .001”) are recalculated the real p value is found to be extremely and implausibly small, e.g., *p* < 10-40. This has labelled a “smaller than a lowest threshold” (STALT) inconsistency and can be used to detect implausibly extreme p values Heathers & Meyerowitz-Katz, 2024).

### 

### Are the means and variances of integer data impossible?

* Depending on sample size and number of decimal places to which summary stats were reported to, the tools GRIM (means) and GRIMMER (SDs) can be used to check if these are possible given a) the sample size and b) the amount of averaging already done (e.g., if data represent the mean of multiple items).
  + See [Brown and Heathers (2017)](https://journals.sagepub.com/doi/abs/10.1177/1948550616673876) on GRIM, [Allard (2018)](https://aurelienallard.netlify.app/post/anaytic-grimmer-possibility-standard-deviations/) on GRIMMER
  + Lukas Jung’s tool <https://errors.shinyapps.io/scrutiny/>
  + Note: There are other simpler web apps for GRIM and GRIMMER too, but they don’t let you save results and they have some hidden settings/assumptions that make it easier to misuse them.
* Regardless of sample size, when truncated data (i.e., has logical min and max, such as 1 and 7 on a 1-7 Likert scale) sets of M and SD (and N, semi-optionally) can be checked for inconsistencies. Not all combinations of mean and SD are possible.
  + See Ian’s tool TIDES: <https://errors.shinyapps.io/TIDES/>

### 

### Are there errors or inconsistencies in statistical analyses?

* Need to make sure people don’t just try to reproduce the result based on rounded summary data – don’t expect to match.
* Instead, need to check whether reported p-values are consistent with the rounded summary data – what can we direct people to use here (Nick Brown’s spreadsheet, or make something new? Web app would be useful)
  + Ian and Lukas have R packages and tools for this in early development. These will allow users to recalculate various statistics (p values, effect sizes) from reported test statistics or summary statistics while taking rounding into account.
* The consistency between reported *p* values, test statistics and df can be compared for various common tests (t tests, ANOVAs, Chi-squared tests) using the tool statcheck (<https://statcheck.io>), which also has an R package.

### 

### Are any other contradictions implied by the data?

* This item is intended to represent both a) a larger number of subchecks that may be relevant to only some articles, and b) to provide an ‘other’ category under which researchers can list other checks they have done.
* Specific subchecks include:
  + Comparing sample size (N), mean, and Standard Deviation between a) the overall sample and b) the combined subgroups.
    - See Ian’s tool ANCHOR: <https://errors.shinyapps.io/ANCHOR/>
  + Checks of range vs other results. For example, a 5-item scale that uses 1-7 Likert response options must have a sum score between 5 and 35. If the reported mean is outside the range [5,35], this represents an inconsistency.
  + Checking whether correlation Pearson’s *r* correlation tables are mathematically possible (i.e., if they are positive definite)
    - See Ian’s tool PORT: <https://errors.shinyapps.io/PORT/>

### 

### Are there inconsistencies in descriptions of methods and results across publications describing the study?

* Check for consistency between multiple publications by the authors of the study referring to the study – check results and methods (including reported dates etc). For example, if a protocol paper is published and then one or more articles with empirical results from the study, details of the study should agree between them.

What to do with overall assessment?

* Should be in line with Cochrane policy on managing potentially problematic studies.
* If any “serious concerns”: put study in “awaiting assessment” category and exclude from review and meta-analysis.
* If any “some concerns”: conduct a sensitivity analysis with and without these studies.

# Proposals for other types of checks more relevant to psychology studies

## Inspecting measures in the paper

### Does the article use ad hoc measures?

* Were the (key) measures (e.g., the ones used to collect data used to test the central hypotheses) created for this study, i.e, have they not been used in any previous studies nor were they developed and validated in a separate sample to that used to test hypotheses?
  + What is the evidence that the claims do not rest on a nominative assumption?
  + If ad hoc, what is the evidence against Schmeasurement/flexibility/overfitting?

### Do the measures involve self-reports about behavior?

* If yes, is there any evidence that self-reports about this behavior are adequately associated with that overt behavior?

### Are standardized effect sizes plausible given the reliability of the measures used in them?

* Correlations between X and Y are limited by the reliability of X and Y, e.g., if the reliability of both X and Y is Cronbach’s alpha = .70 then the max observable correlation is not r = 1.0 but r = .70. This gets more complicated when the two measures have different reliability, but it still holds that reliability places a bound on observed association.
* Likewise, Cohen’s d can be de-attenuated for the correlation of the measure, and it can be worth asking whether the magnitude of not only Cohen’s d but also de-attentuated Cohen’s d is plausible. However, there is far less normative data on de-attentuated Cohen’s d.
* DOCC: Dis-attenuate Correlations and Cohen’s d <https://errors.shinyapps.io/docc/>

### Are changes across time supported with evidence that measurement did not change over that period?

* Determining trends over time requires a degree of confidence that changes are due to variation in the phenomenon and not merely changes in the way that the variable was measured.
  + E.g., maternal mortality rate, but also depression pre-post therapy.