Quality & Quantity (2024) 58:2929–2955
https://doi.org/10.1007/s11135-023-01758-w

# Tipping the analytical scales, investigating the use of frequentist equivalence analyses in psychology: a scoping review

**Alex D. Marshall**[1] · **Stefano Occhipinti**[2] · **Natalie J. Loxton**[1]

Accepted: 15 September 2023 / Published online: 30 November 2023

## Abstract

Psychological researchers may be interested in demonstrating that sets of scores are equivalent, as opposed to different. If this is true, use of equivalence analyses (equivalence and non-inferiority testing) are appropriate. However, the use of such tests has been found to be inconsistent and incorrect in other research fields (Lange and Freitag 2005). This study aimed to review the use of equivalence analyses in the psychological literature to identify issues in the selection, application, and execution of these tests. To achieve this a systematic search through several databases was conducted to identify psychological research from 1999 to the 2020 that utilized equivalence analyses. Test selection, choice of equivalence margin, equivalence margin justification and motivation, and data assessment practices for 122 studies were examined. The findings indicate wide variability in the reporting of equivalence analyses. Results suggest there is a lack of agreement amongst researchers as to what constitutes a meaningless difference. Additionally, explications of this meaninglessness (i.e., justifications of equivalence margins) are often vague, inconsistent, or inappropriate. This scoping review indicates that the proficiency of use of these statistical approaches is low in psychology. Authors should be motivated to explicate all aspects of their selected equivalence analysis and demonstrate careful consideration has been afforded to the equivalence margin specification with a clear justification. Additionally, there is also a burden of responsibility on journals and reviewers to identify sub-par reporting habits and request refinement in the communication of statistical protocols in peer-reviewed research.

**Keywords** Equivalence analyses · Equivalence test · Non-inferiority · Scoping review · Frequentist analysis

✉ Alex D. Marshall
alex.marshall@griffithuni.edu.au

[1] School of Applied Psychology, Griffith University, Mt Gravatt, Brisbane, QLD 4122, Australia

[2] International Research Centre for the Advancement of Health Communication, Department of English and Communication, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

# 1 Introduction

Increasingly, psychological researchers are faced with questions such as whether or not a new therapy is just as good, or no worse than, a pre-existing therapy (Rogers et al. 1993). However, traditional statistical analyses (based on the classical Fisher–Neyman–Pearson paradigm) are predicated on the assumption that two groups may be different, with a null hypothesis representing no difference or no effect. Such tests, referred to as superiority testing, could not provide answers to questions regarding whether or not groups are the same as these statistical predictions would map onto the null hypothesis and the latter can only be rejected. Rather, such research questions require procedures grouped under the rubric of equivalence analysis (Schuirmann 1987). Equivalence analyses provide specialised procedures that enable researchers to test if groups of scores come from the same population—the logic of which runs opposite to the foundations of superiority testing. Furthermore, equivalence analyses necessitate several critical judgments, such as the specification of an equivalence margin that are absent in superiority testing paradigms. However, existing systematic reviews in other disciplines suggest that researchers using equivalence analyses are inconsistent in the selection, execution, and reporting of such analyses. Further, these reviews indicate researchers frequently fail to follow recommendations from educational material, which is often inconsistent or vague in or of itself (Althunian et al. 2017; Lange and Freitag 2005; Rehal et al. 2016; Wangge et al. 2010). Additionally, many existing reviews are situated within medical research fields (Althunian et al. 2017; Lange and Freitag 2005; Rehal et al. 2016; Wangge et al. 2010). This is problematic because issues faced in medical research fields are not necessarily relevant to psychological research. For example, in many medical fields, stipulation of the equivalence margin, the range within which confidence intervals are seen to demonstrate equivalence (further explanation to follow), may be strongly determined by external factors, such as the Food & Drug Administration (e.g., FDA 1992). By contrast, no such governance exists within psychology, and researchers stipulate their margins independently. Therefore, findings and recommendations regarding the reporting conventions of medical researchers may not be relevant to psychological researchers. Accordingly, the present paper reports a scoping review into the selection, application, and execution of equivalence analyses (i.e., equivalence tests and non-inferiority tests) in psychological research.

## 1.1 Existing reviews

Reviews examining the use of equivalence analyses have focussed heavily upon aspects of the statistical design that are relevant for establishing equivalence. For example, Lange and Freitag (2005) reviewed equivalence analyses in medical research. Their findings focus upon quantification of the equivalence margins, justifications for such margins, and the sample selection (i.e., intent-to-treat versus per-protocol). Their review found that the selection, application, and reporting of such analyses is largely inconsistent and heterogenous (Lange and Freitag 2005). However, since this review was published the medical field has seen the establishment of guidelines and recommendations for conducting equivalence trials (Piaggio et al. 2006, 2012). Thusly, it follows that the quality of reporting of equivalence analyses should now have increased considerably.

More recent reviews, for example, by Wangge et al. (2010), Rehal et al. (2016) and Pong et al. (2021) may provide insight into the use of equivalence analyses following the addition of new guidelines (Piaggio et al. 2006, 2012). However, despite the introduction of refined guidelines, these reviews demonstrated that study design and analyses are still inconsistent and of poor quality, suggesting that equivalence analyses are still not well understood (Althunian et al. 2017; Pong et al. 2021; Rehal et al. 2016; Wangge et al. 2010). Additionally, these reviews did not address psychological research, but rather medical research. Specifically these reviews tend to examine clinical trials, and outcomes such as bioequivalence or mortality rates of treatment (Pong et al. 2021). While equivalence analyses are utilised in clinical psychological trials in psychology, such tests are often applied to a range of non-clinical research questions. For example, Lewis et al. (2009) and Epstein et al. (2001) conducted equivalence analyses on measurement approaches (e.g., pen-and-paper versus internet-based measurement). Beringer and Ball (2009) provide an additional example, where the authors applied equivalence analyses to understanding the interpretation of in-flight heads-up-displays for pilots during flights. These examples highlight that the nature of psychological research is extremely varied and taking a narrow approach by only examining clinical research would not provide an accurate quantification of the extant literature using equivalence analyses.

Given that medicine research differs heavily from psychological research, and that the focus of existing reviews has largely centred on clinical trials, the relative utility of these existing reviews may be limited in psychology. For these reasons, a review of psychological literature would allow for the identification of strengths and limitations in the selection, application, and execution of equivalence analyses in the field. Further, quantification of these strengths and limitations would allow for the development of targeted guidelines that would help to increase the quality of such analyses in psychology.

Finally, existing reviews primarily examine a single effect or phenomenon and are less concerned with a broad mapping of the current state of the extant literature. A scoping review differs from existing reviews, which are frequently systematic in nature (e.g., Althunian et al. 2017; Lange and Freitag 2005; Pong et al. 2021). Given that psychological literature varies widely between subdisciplines, and the primary issues to be addressed are not specific to a subdiscipline or study design (e.g., clinical trials) but rather require synthesis of broader research factors, a scoping review is appropriate (Arksey and O'Malley 2005). To this end, this present study presents a scoping review into the selection, execution, and reporting of equivalence analyses within the psychological literature.

### 1.2 Equivalence analyses

Equivalence analyses were first developed in areas such as pharmacology (e.g., Schuirmann 1987) to address the research question that two treatments (e.g., one being the more expensive or having more adverse impacts) do not differ in their therapeutic effect. In the case of non-inferiority analyses, the test responds to the research question that one treatment is *not worse than* another (Schumi and Wittes 2011). Importantly, while equivalence and non-inferiority are two separate approaches to two different research questions, the way in which they are conducted is similar. This is especially true for the establishment of the equivalence margin. As such, this scoping review will refer to equivalence analyses to encompass both equivalence tests and non-inferiority tests. See Leichsenring et al. (2018) for more specific and nuanced descriptions of these tests with contrasts to typical superiority testing.

The application of typical null hypothesis significance testing (NHST) approaches to research questions that are primarily concerned with equivalence (or non-inferiority) of scores is erroneous, because research questions of equivalence would assert the null hypothesis, resulting in unfalsifiable predictions. From statistical and logical perspectives, one cannot establish equivalence with non-significant superiority tests (Rogers et al. 1993; Schuirmann 1987). A failure to reject the null hypothesis in superiority testing would only allow a researcher to argue that the groups of interest were not different or that the test lacked sufficient statistical power to reject the null (Lakens 2017). A lack of an effect does not quantify equivalence because superiority testing does not establish criteria with which to define *meaningless* differences. By contrast, when employing equivalence analyses researchers construct a set of falsifiable hypotheses that together provide evidence of equivalence, rather than no evidence of superiority (Lakens 2017; Rogers et al. 1993). Generally, equivalence analyses have been applied to a range of research areas including medicine, communications, and physiotherapy. One of the more recent applications of such analyses has been to psychology.

### 1.3 Equivalence analyses in psychology

Examples of research where demonstration of equivalence is the primary goal are plentiful within psychological research. For example, the aim may be to demonstrate that a shorter, more intense cognitive-behavioural therapy (CBT) protocol is equally efficacious as a standard CBT protocol offered to individuals with depression. The researcher's question is whether the outcomes of the intense, treatment protocol are equal to the current gold-standard approach (or in the case of non-inferiority testing, that the outcomes of the intense treatment protocol are *not worse than* the current gold-standard). Importantly, equivalence analyses involve several key deviations from superiority testing that researchers must be familiar with if they are to appropriately apply these paradigms. One primary consideration to be made concerns the null and alternative hypotheses relevant to equivalence analyses. In traditional superiority testing with two independent samples, the null hypothesis posits that the two samples' scores arise from the same population and any differences in observed means are due to random error: that is, the population means of the samples are the same. The alternative hypothesis posits the inverse. In equivalence analyses, these hypotheses are reversed. The null hypothesis posits that the samples are not from the same population (i.e., not equivalent). Importantly, this ensures that a falsifiable prediction can be generated and enables logically correct evidence to be drawn from the tests to demonstrate an effect. Finally, like superiority testing, equivalence analyses can be used to investigate several statistics. One familiar approach to psychological researchers would be the comparison of two independent group means. However, equivalence analyses can also be applied to investigate the equivalence of dependent (within-subject) group means, correlation coefficients, and meta-analyses (Lakens 2017). Equally, the mechanisms that underlie these analyses involve the specification of the equivalence margin, a unique challenge of equivalence analyses procedures.

### 1.4 Equivalence margin

In practical terms, equivalence analyses also need to account for random chance; even when observed samples do arise from the same population, the means may differ only because of random variation. The equivalence margin (common nomenclature includes equivalence interval, equivalence bounds, and margin of equivalence) is an important component of such analyses that addresses this issue. This margin represents an area within which the means of two sets of scores can be different yet still be considered equal (Rogers et al. 1993). An example for visualisation of such an analysis is presented in Fig. 1. As can be seen, conducting an equivalence analysis is akin to constructing a confidence interval around the mean difference of two sets of scores. If the confidence interval is entirely encapsulated within the equivalence margin, it can be said that these sets of scores are equivalent (Lakens 2017). As seen in other frequentist inferential statistics, researchers can use the *p* values of the Two One-Sided Test (TOST) procedure to examine equivalence of the groups. The TOST procedure will involve two simultaneous tests, one of the upper portion of the margin and one of the lower portion of the margin. As such, there are two *p* values to consider; if both values are below the nominal alpha (typically equal to 0.05) then one can infer that the groups are equivalent. However, if just one of the *p* values is above the nominal alpha (e.g., exceeds 0.05) one can infer that the groups are non-equivalent. In Fig. 1, both tests are statistically significant at an alpha of 0.05, visually this is represented as the mean difference (black square) being entirely encapsulated by the equivalence margin (vertical dashed lines at 2, − 2 on the x-axis).

Careful consideration should be afforded to the equivalence margin because an incorrectly specified margin results in a meaningless equivalence test. In order to maintain Type I error rate (i.e., a false positive, demonstrating equivalence of groups despite the population-level effect being non-equivalent) at a desired level, the margin must be set a priori rather than after having seen the data (Schuirmann 1987). If the margin is established post-hoc, it would be difficult to ensure that no biases have confounded the appropriateness of the equivalence margin (Rogers et al. 1993). Further, the margin needs to be sufficiently specific to reflect true meaninglessness.

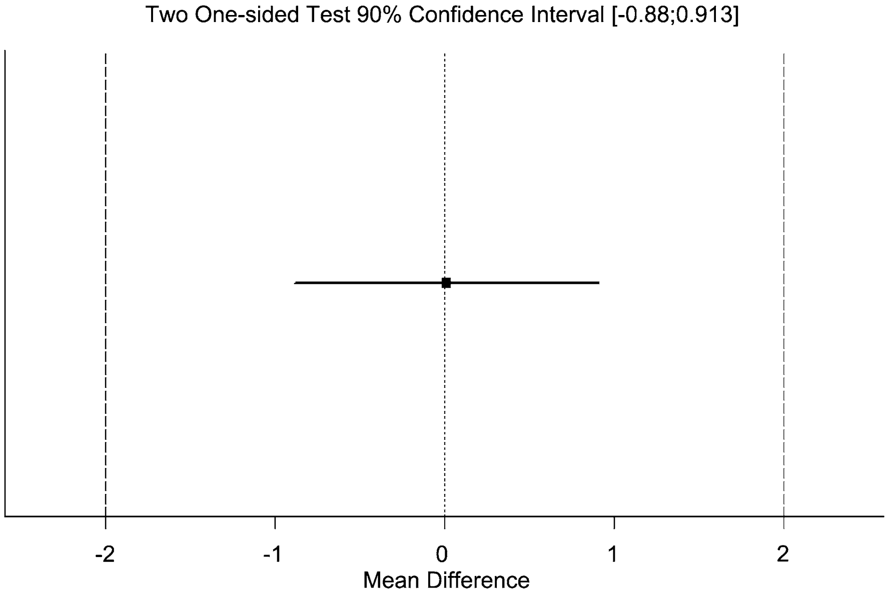

**Fig. 1** Example Visualisation of an Equivalence Analysis, Specifically the Two One-sided Test Procedure. *Note.* Vertical dashed lines indicate the equivalence margin. The vertical perforated line indicates a mean difference equal to 0. The square in the centre indicates the mean difference between two groups. The horizontal black line indicates a 90% Confidence Interval around the mean

The quantification of meaninglessness is a critical aspect of equivalence analyses. Margin specification involves balancing several aspects of research including study feasibility limitations (e.g., recruitment resources). However, these practical considerations are not as important in the conduction of equivalence analyses as the theoretical issues researchers face in quantifying meaninglessness. Consequently, more practical issues should not be the primary motivator for specifying an equivalence margin, and it is the theoretical issues that should be considered more carefully. In psychology it is unlikely that clear, explicit, and direct information is available from theoretical perspectives that quantify meaninglessness. Psychological researchers face a unique challenge in that, in the face of ambiguity surrounding theoretical quantifications, they are tasked with establishing the margin independently and without clear guidance. Indeed, research suggests the use of effect size benchmarks (e.g., $d=0.3$) in the absence of clear directives for theoretical meaninglessness may be appropriate (Lakens 2017). However, it is unclear if these benchmarks are appropriate for psychological research. For example, the use of effect size benchmarks is still arbitrary (Cohen 1988). Additionally, together with issues associated with quantifying meaninglessness, the equivalence margin must have an appropriate width such that study feasibility (e.g., the need to gather an exorbitantly large sample) does not prohibit study completion (Cribbie et al. 2004; Lange and Freitag 2005; Walker and Nowacki 2011). If the equivalence margin is too narrow, it may be impossible to demonstrate equivalence of scores even when the sets of scores are drawn from the same population (without collecting data from the whole population). Conversely, if the equivalence margin is too wide the differences within the margin would not accurately reflect practical or theoretical meaninglessness (Walker and Nowacki 2011).

These issues suggest that establishing the margin requires significant deliberation. In some instances, margin specification may draw upon several resources, including but not limited to, expert researcher or clinician consultation, existing literature, and pilot data (Lange and Freitag 2005). However, this may not always be possible (e.g., if there are no resources to draw upon), and instead, researchers may establish their margin based only on their perception of a meaningless difference between the sets of scores. In medical research external factors commonly govern the equivalence (or non-inferiority) margin. For example, the U.S Food and Drug Administration (FDA) recommends establishing a margin at 20% of the reference group mean for drug trials in areas such as pharmacokinetics (FDA 1992). No such governance exists in psychological research. In the absence of externally stipulated standards (e.g., what the FDA specifies as an appropriate equivalence margin), psychological researchers rely much more on researchers approximating meaningless. In psychology, the margin is intrinsically related to cognition and behaviour, and measurement of such phenomena involve considerable error relative to areas from which this statistical approach originated (e.g., cell metabolization, a construct measured with high accuracy; Alavijeh et al. 2005). Consequently, psychological researchers must establish a margin that is sufficiently narrow to ensure equivalence of groups is meaningful, while balancing the unduly effects of poorer measurement tools (relative to other research areas, such as pharmacology). Because of this, establishing a margin that truly reflects meaninglessness is not a trivial process; however, given the relative recency of equivalence analyses in psychological literature, a strong set of discipline specific norms regarding margin stipulation is unlikely to have developed.

Additionally, to communicate that the specified margin reflects meaningless, authors should provide a justification for the specification of such a margin (Althunian et al. 2017; Lange and Freitag 2005; Rehal et al. 2016); however, as is evidenced by previous reviews, justifications for margins are often vague, inconsistent, or missing entirely (Althunian et al.

2017; Lange and Freitag 2005; Rehal et al. 2016). This results in an equivalence analysis that is difficult to interpret and severely limits the generalisability of inferences drawn from such tests (Cribbie et al. 2004; Lange and Freitag 2005). Further, researchers should be aware of any data related complications (e.g., assumption violations) that may affect the validity of their statistical inferences.

### 1.5 Assumption checking and outlier management

Similar to other tests from the general linear model, specific equivalence analyses are subject to a set of assumptions (Schuirmann 1987). A failure to meet the assumptions associated with a test may result in biased coefficients leading to an increase in Type I error rate or reduced statistical power (Glass et al. 1972; Srivastava 1959), because certain analyses assume that sample(s) and population(s) meet specific criteria regarding their structure (Nimon 2012). Despite the demonstrated need to verify that certain assumptions are at least not violated too severely, Hoekstra et al. (2012) suggested that researchers are misinformed with respects to the relevant assumptions of popular tests (e.g., *t* tests) and how to check these assumptions. Given the relative infancy of equivalence analyses in psychology, and that typical statistical training does not address equivalence and non-inferiority tests, it is unclear how familiar researchers are with assumptions relevant to this novel approach.

In the case of the TOST procedure for example, a set of assumptions exist that should be met prior to conducting and interpreting the test results. It is assumed that the data is normally distributed, the variance across groups is homogenous in nature and the sample size is relatively large (at least 30 observations; Schuirmann 1987; Witte and Witte 2017). Additionally, tests such as the TOST procedure are often subject to the effects of outliers (i.e., extremely influential scores; Bakker and Wicherts 2014). Statistical outliers may bias or distort otherwise accurate statistical models, and it is generally accepted that outlier detection techniques be used prior to conducting the chosen analysis (Bakker and Wicherts 2014; Smiti 2020). Despite the availability of information surrounding assumptions and the effects of outliers, the predominant educational material neglects aspects of these issues, or entirely neglects the topic (Lakens 2017; Lakens et al. 2018; Leichsenring et al. 2018; Rogers et al. 1993). Additionally, existing reviews on the use of equivalence analyses in other research areas have neglected assumption checking and outlier identification protocols reported in studies (Althunian et al. 2017; Lange and Freitag 2005; Wangge et al. 2010). Given these factors may bias statistical tests, this present scoping review aims to quantify the extent and nature of reporting on assumption checking and outlier identification processes in psychological research using equivalence analyses. Another specific issue researchers face with equivalence analyses in clinical trials is sample selection.

### 1.6 Selecting the analysis samples

One specific issue clinical trials face using equivalence analyses is sample selection for analysis—intention-to-treat (ITT) versus per-protocol (PP). This issue has been debated for several years in medical research fields because clinical trials form a large basis of the research in medicine (Piaggio et al. 2006, 2012). Given this, it is reasonable to expect psychological researchers to first turn to the recommendations and regulations established in fields such as pharmacodynamics and bioequivalence trials when selecting their analytical sample. For superiority testing, the ITT sample (participants analysed based on their treatment allocation, regardless of treatment adherence) is generally preferable. The ITT sample is believed to underestimate the true effect of a treatment in a superiority setting, and therefore it reduces the risk of committing a Type I error (i.e., falsely finding a difference between two measures of an outcome variable; D'Agostino et al. 2003). In such an instance, the treatment effect is underestimated because all individuals involved in a treatment arm are involved in the final analysis, and non-adherent, or less adherent participants diminish the efficacy of the treatment (Gøtzsche 2006). As a result, the ITT sample will tend to demonstrate reduced or no differences in treatment arms, thereby assisting to safeguard against false positives in superiority settings. This poses an issue in equivalence (or non-inferiority) designs, because ITT samples would tend to lead to the desirable outcome (diminished differences in treatment arms; Gøtzsche 2006; Gupta 2011). Contrastingly, the PP sample (i.e., only participants who are adherent to treatment are included in the final analysis), in a superiority setting is believed to reflect more accurately the differences between two treatment arms. As only treatment-adherent individuals are retained in the final analysis, the PP sample is believed to provide a more accurate estimate of the *true* treatment effect. Both the ITT and PP samples offer estimates of different treatment effects, and for equivalence analyses, sample selection has been debated several times (Gøtzsche 2006; Matsuyama 2010; Piaggio et al. 2006, 2012).

More recently, recommendations from existing literature have encouraged reporting of both the ITT and PP samples (Piaggio et al. 2006, 2012). This is recommended because the ITT and PP samples provide different information that can be relevant to the research question being investigated. ITT samples will tend to provide an estimate of the overall treatment effect (e.g., CBT intervention). This occurs because not all individuals from the population will undertake the treatment for its full course, as prescribed. Contrastingly, a PP sample tends to provide an estimate of the overall effect for a full course of treatment, as prescribed (e.g., the effect of "perfect" treatment adherence). While the information gathered is different, both aid in comparing the relative efficacy of treatment and as such, analyses on both the ITT and PP samples should be reported (Piaggio et al. 2012). Although there are benefits to reporting on both samples, existing literature reviews show that authors tend to reject the guidelines available and frequently only report analyses of one sample (Kay 2014; Lange & Freitag 2005; Le Henanff et al. 2006). Given the discrepancies between existing guidelines and issues in reporting the ITT versus the PP analysis samples in the medical literature, this present scoping review aims to quantify the sample selection and reporting processes for clinical trials in psychological research where sample selection is relevant.

### 1.7 Aims of the present review

In sum, existing reviews of the use of equivalence analyses demonstrate inconsistencies in the selection, application, and execution of such tests. However, these reviews have limited application to psychological literature because psychology has been largely overlooked. As an extension of this, the recommendations laid forth by existing reviews may have limited utility for psychological researchers looking to implement equivalence analyses. If psychology as a science is to develop, particularly with respects to the use of equivalence analyses, the literature needs to be examined for inconsistencies so that issues may be addressed. To this end, these inconsistencies may provide a clear direction for future research that will lead to increased proficiency in the use of equivalence analyses.

Given the current state of the literature, the present study presents the findings of a scoping review into the use of equivalence analyses in psychological research. This review aims to ascertain several key characteristics of the literature. Given that equivalence analyses depend on the selection of a statistical model for determining equivalence (or non-inferiority), and on the specification of an equivalence margin, the first and second aims of this review are to quantify the most used tests of equivalence and the most reported equivalence margins, respectively. Further, given the importance of the equivalence margin, the third aim is to quantify the extent to which researchers describe and justify their equivalence margin specification process. The final aim of this review focuses on the reporting of assumptions, outlier identification and management and trial-specific sample selection. Given that inferences based upon equivalence analyses are subject to bias due to data-related issues, this present review aims to quantify the extent to which assumptions and outliers checking are reported. Additionally, for clinical trials, this present review aims to quantify the nature of sample selection.

## 2 Method

### 2.1 Transparency and openness

This review follows the reporting guidelines presented in the Preferred Reporting Items for Systematic and Meta-analyses Protocols Extension for Scoping Reviews (PRISMA; Tricco et al. 2018). Additionally, this review follows the recommendations for conduction and reporting of scoping reviews as advocated by Arksey and O'Malley (2005). Further to this, how the data was identified, evaluated for inclusion in the review and charted are made explicit. The charted data and associated notes are available at https://osf.io/v7j5z/?view_only=8de42816bea64215b4500392f2de8922.

### 2.2 Eligibility criteria

This review included only peer-reviewed journal articles published between 1999 and 2021 that conducted an equivalence analysis as a component of the statistical protocol. The timeframe was expected to yield sufficiently representative results. The articles had to conduct an equivalence analysis to answer a research question primarily concerned with psychological phenomena. This is to say that the included studies investigated mind and/or behaviour. The search strategy was restricted to include papers only published in English. Articles that did not meet the above inclusion criteria were excluded as they were deemed irrelevant or out of the scope of the review.

### 2.3 Literature search strategy

The literature search was conducted primarily by one member of the research team. The literature search was limited to the databases Embase, PsycINFO, PubMed and, Scopus. Grey literature was retrieved in two ways: (1) manual searches completed with the Griffith University database and Google Scholar and (2) reference scanning of all articles included in the full-text review. The search strategy began by constructing a comprehensive search string and conducting a search in each of the databases. After narrowing the initial body of studies from 11,260 to 101 for full-text review (via title and abstract, and keyword scanning), two manual searches were completed in the Griffith University database and Google Scholar. Following this, the reference list of each article included in the full-text review was scanned for any relevant studies. The function of the manual searches and reference list scanning was to identify and gather any articles missed by the initial searches that may have been relevant to the review. The manual searches in conjunction with the reference list scanning facilitated the identification of an additional 12 relevant papers. Approximately 12 months from the initial search for this review, a second follow-up search was conducted to ensure papers published since the first search were considered for inclusion in this review. This process involved repeating the database searches above (however, there was no follow-up manual searches or reference list scanning). This resulted in an additional nine papers being identified. As a result, this present review found a total of 122 papers that met all inclusion criteria. See Fig. 2 for an overview of the literature search and screening process.

To ensure that the papers omitted and included for this review were appropriate an inter-rater reliability task was completed. This process involved randomly sampling 20 papers that were omitted from full-text review and 20 papers that were included in the full-text review. Another member of the research team reviewed each paper and decided to omit or retain the paper from the review. When comparing the two raters' judgments, there were no discrepancies as to the omission or retention of papers, with 100% agreement across all 40 articles.

## 2.4 Search string

The search string for this review had three components. First, it included variations of the term "equivalence analysis", to capture the various ways in which equivalence analyses can be described. Second, the search string included variations of the term "non-inferiority analysis". Finally, the search string included a Boolean function to eliminate two research

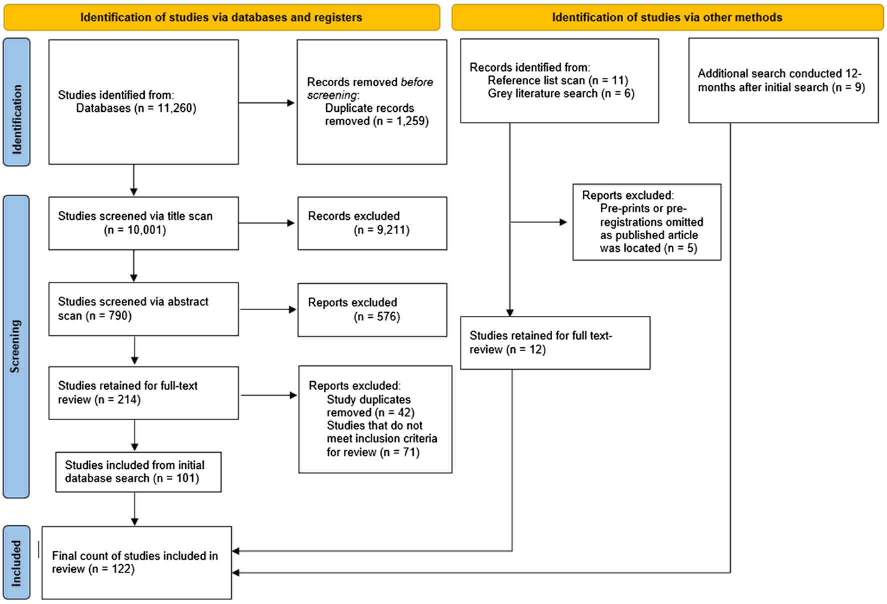
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Studies identified from:
Databases (n = 11,260)
Records removed *before screening*:
Duplicate records removed (n = 1,259)
Records identified from:
Reference list scan (n = 11)
Grey literature search (n = 6)
Additional search conducted 12-months after initial search (n = 9)
Screening
Studies screened via title scan (n = 10,001)
Records excluded (n = 9,211)
Reports excluded:
Pre-prints or pre-registrations omitted as published article was located (n = 5)
Studies screened via abstract scan (n = 790)
Reports excluded (n = 576)
Studies retained for full text-review (n = 12)
Studies retained for full-text review (n = 214)
Reports excluded:
Study duplicates removed (n = 42)
Studies that do not meet inclusion criteria for review (n = 71)
Studies included from initial database search (n = 101)
Included
Final count of studies included in review (n = 122)

**Fig. 2** CONSORT diagram of the literature search and screening process

areas (i.e., pharmacokinetics and pharmacodynamics). These research fields are populated by equivalence analyses to a far greater degree than almost any other field in that the use of bioequivalence trials and equivalence analyses are pertinent to many research questions in these fields. Retaining these fields, the search resulted in an almost insurmountable number of papers that would have been infeasible to scan. Further, scans suggested that these fields rarely involved any psychological phenomena as primary outcomes, so including these fields served little function in completing the present review. Titles, abstracts, and keywords were screened using the following search string:

> Equivalen* test*" OR "equivalen* analy*" OR "equivalen* trial*" OR "non-inferiority analy*" OR "non-inferiority test*" OR "noninferiority analy*" OR "noninferiority test*" OR "non-inferiority trial*" OR "noninferiority trial*" NOT "pharmacokinet*" OR "pharmacodynam*"

### 2.5 Data charting process and data items and synthesis of results

The data charting process for this review involved constructing a large table that housed key information required to answer the research questions presented above. This table was constructed in Microsoft Word and can be found in the supplementary material.

Results were synthesized with narrative form and tables. Narrative form involved extracting the main themes relevant to answering the research questions and connecting these major themes across the body of literature. Creation and population of the table allowed for each datapoint to be tracked, facilitating the synthesized analysis. The charting process and data analysis was primarily conducted by one member of the research team.

## 3 Results

Refer to the supplementary material for the charted raw data analysed in this scoping review (available at https://osf.io/v7j5z/?view_only=8de42816bea64215b4500392f2de892).

### 3.1 Selection of the equivalence analysis

The first aim of this review was to ascertain the tests frequently used to specify equivalence. Broadly, there are two categories that capture almost all specified tests in this scoping review (1) equivalence tests and (2) non-inferiority tests. There exists a third category that encapsulates all other approaches, accounting for only a very small subset of papers in this review.

#### 3.1.1 Equivalence testing

Of the 122 papers examined, 42 utilized equivalence testing in their analysis. The most popular approach in this review was Schuirmann's (1987) TOST procedure, with 15 papers specifying this approach. The use of inferential confidence intervals as detailed by Westlake (1972) and Rogers et al. (1993), were the next most utilized approaches. These approaches involve specification of the CIs, with 13 papers using 90% CIs and 10 papers using 95% CIs. A further three papers utilized Wellek's critical constant approach (Fals-Stewart et al.

2005; Fals-Stewart and Lam 2008; Schmitz et al. 2001). And finally, a single paper used Wellek's goodness-of-fit approach to assess the equivalence of modelling (Gagnon et al. 2016).

### 3.1.2 Non-inferiority testing

Of the 122 papers examined, 76 used non-inferiority testing and inferential confidence intervals comprise the majority of tests in some form. The most common was the use of a one-sided 95% CI, with 33 papers using this approach. A further 25 papers reported the use of a two-sided 95% CI approach. The remaining tests appear at reduced frequencies, with two papers reporting the use of a one-sided 90% CI approach and five papers reporting the use of a two-sided 90% CI approach. A further three papers reported the use of a one-sided 97.5% CI, and a single paper reported the use of a one-sided 98.75% CI approach. Conversely, there are several papers that fail to adequately report the test specified: in 7 papers authors reported only parts of the tests utilized. For example, Dirkse et al. (2020) reported the use of a non-inferiority test but did not report the CI for the approach. This lack of information varies, with some papers failing to report on the selected CIs, the direction of their test (one- or two-sided), and at times failing to specify a statistical test at all.

### 3.1.3 Other approaches

Of the remaining five papers identified in this review, there were various approaches used. Beck et al. (2018) reported the use of an ANCOVA to test for non-inferiority. It is unclear how specifically the ANCOVA was used as a test of non-inferiority. de Zwaan et al. (2012) reported the use of a two-sample *t*-test to test for equivalence. Finally, two papers did not state their selected equivalence analysis (Mathiasen et al. 2016; Romijn et al. 2015).

## 3.2 Selection of the equivalence margin

The second aim of this review was to quantify the most frequently specified equivalence margins. To achieve this aim, margins were categorized by units of expression: percentages/proportions, raw scale scores, standard deviations, and effect sizes. Below is the presentation of such findings and reporting on the most specified equivalence margins (refer to Fig. 3).

Of the 122 papers reviewed, 32 expressed their margin in proportions or percentages of the mean. This occurred to various degrees; however, the most popular approach was the use of 20% of the reference group mean, with 12 authors using this approach. A further eight papers specified a margin equal to 15% of the reference group mean. Three papers used 10% of the reference group mean. Two papers used 25% of the reference group mean, and two papers used 5% of the reference group mean. The remaining three papers used proportions equal to 50%, 17.85% and 15.95% (grose Deters et al. 2014; Hofmann et al. 2015; Malinvaud et al. 2016).

Sixty papers expressed the equivalence margin in raw scale scores. This category houses the largest number of papers reviewed in the present study and is also the most variable in nature, in that no discernible patterns could be identified by standardizing the raw scales scores presented. Compounding this issue was the identification of instances where papers from specific sub-disciplines reported different equivalence margins for the same constructs and outcome variables (e.g., Driessen et al. 2017; Ly et al. 2015). This issue also

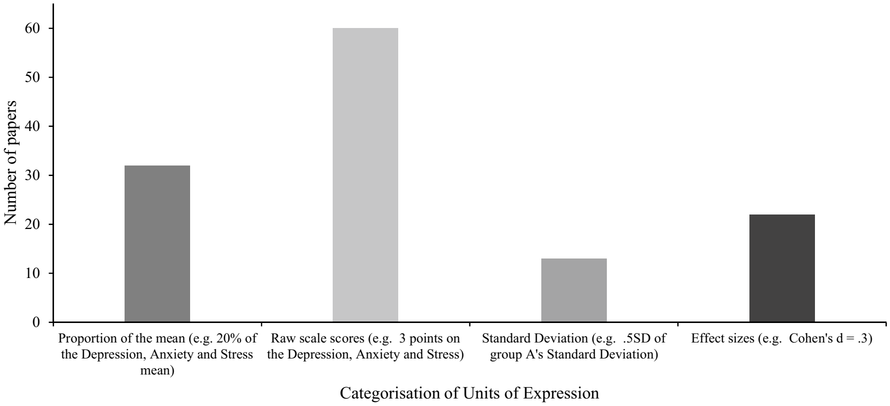

**Fig. 3** Frequency of unit of expression for the equivalence margin

exists amongst papers that examined the same constructs in the same population (e.g., Acierno et al. 2017; Mathersul et al. 2019).

Thirteen papers expressed the equivalence margin in standard deviations. Ten papers used half a standard deviation (i.e., 0.5SD) and the remaining three papers used various levels. Gagnon et al. (2016) reported an equivalence margin equal to a standard deviation of 1, Ball et al. (2013) utilised one-third, two-thirds and a full standard deviation (and interpret their analyses with all 3 specifications), and Norton and Barrera (2012) express their margin using 0.6 of a standard deviation. There was little consistency between papers with respects to the standard deviation selection. For example, Gagnon et al. (2016) used the standard deviation estimate from a population (the authors had access to this data before conducting the study). However, Liu et al. (2019) used half of the pooled standard deviation between the examined groups. Finally, Gross et al. (2019) do not explicate which standard deviation is used (e.g., it is unclear if the margin is based on the standard deviation of the reference/treatment group, or if it was the pooled standard deviation from both groups). In many cases, authors do not distinguish what standard deviation is used, and this reduces clarity regarding margin specification.

Twenty-two papers expressed the equivalence margin in effect size. Nine papers used an effect equal to Cohen's $d = 0.3$ (e.g., $-0.3$ to $+0.3$ representing the equivalence margin; Charig et al. 2020; Forand et al. 2019; Goldstein et al. 2020). Seven papers reported a margin equal to $d = 0.2$ and five papers used $d = 0.5$ (e.g., Dunn et al. 2019; Romijn et al. 2015; Sloan et al. 2018). Yeung et al. (2018) utilized a margin equal to $d = 0.8$ and $d = 0.4$. Three papers specified their margin with epsilon (ε), however it was not made clear what unit of expression this was (Fals-Stewart et al. 2005; Fals-Stewart and Lam 2008; Gagnon et al. 2016). Lastly, two papers used Hedge's $g = 0.4$ (Hedman et al. 2011, 2014).

### 3.3 Description and justifications for the equivalence margins

The third aim of this present study was to assess the extent to which researchers justified their equivalence margin specification. To analyse this aspect of the data, four categories were defined to aid in understanding these justification (see Table 1). Level of specificity and clarity were the main factors for determining an appropriately comprehensive justification. The four groups represent varying degrees of these factors. Papers in the first category

**Table 1** Categorisation of justifications related to equivalence margin reporting and associated frequencies

| Category | Category description | Category totals | Example papers for references |
|---|---|---|---|
| Clear description and justification | A clear description of the resources accessed when establishing the margin are provided. Concurrently, a justification or argument that emphasises how the differences within the equivalence margin are theoretically or practically meaningless is also provided | 23 | Acierno et al. (2017)<br>Blom et al. (2015)<br>Blumberger et al. (2018)<br>Herbert et al. (2017) |
| Clear description, no justification | A description of the resources accessed when establishing the margin are provided. However, the description is often short and does not include multiple resources. Additionally, no justification or argument is provided emphasising how the differences within the equivalence margin are meaningless | 51 | Alfano (2012)<br>Barlow et al. (2017)<br>Bramoweth et al. (2020)<br>Murray et al. (2018) |
| Vague description, no justification | A vague description of the resources accessed are provided. The description is characterised by a lack of information regarding the relevance of the resources or a poor description of them. The justifications are not present or are too vague to be meaningful | 40 | Andersson et al. (2013)<br>Kuang et al. (2018)<br>Weinstock et al. (2017) |
| Little to no description of justification | Little to no description of the resources are provided. Little to no justification is provided. Descriptions and justifications are characterised by the provision of extremely limited or missing information | 8 | Eschenbeck et al. (2019)<br>Goodman and Israel (2020)<br>Jongen et al. (2017)<br>Malinvaud et al. (2016) |

provide highly specific, clear descriptions and justifications of the equivalence margin. Each subsequent group does this to a lesser extent, reducing specificity and clarity. The last group presented in Table 1 houses papers that provide little to no information regarding the margin, in some cases, the margin itself was not even stated.

### 3.4 Assumption checking and outlier management

The fourth aim of this present review was to examine the extent to which researchers reported on assumption checking, assumption violations and outlier identification and management. Of the 122 papers reviewed, 104 made no comments regarding the assumptions associated with their selected statistical approach or the presence or absence of outliers. Of the remaining 18 papers, 16 make comments surrounding assumption checking, outlier identification and management, or a combination of both. However, these comments are often vague, frequently addressing only one or two issues, and ignoring or disregarding others. Additionally, these comments are not made directly to the equivalence analysis and these issues are reported in a general manner (e.g., Alfano 2012; Beukes et al. 2018). The remaining two papers make specific comments about assumptions and/or outliers as directly related to the equivalence analysis. First, de Zwaan et al. (2012) makes a direct comment about utilising Welch's correction considering an assumption violation concerning homogenous variance between groups. Second, (Bauer et al. 2020) makes specific comments regarding violations to normality assumptions and the presence of outliers, and how these issues are handled during data analysis. These two papers (i.e., Bauer et al. 2020; de Zwaan et al. 2012) are an exception in a body of literature that largely ignores such issues when reporting analyses.

### 3.5 Selecting the analysis samples

The final aim of this present scoping review was to determine the extent and nature of sample selection in clinical trials. To begin, 69 clinical trials were identified in this review, 44 of which reported the use of the ITT sample. An additional 11 papers reported the use of only the PP sample and in the remaining 14 trials, the authors report conducting analyses on both the ITT and PP samples. However, not all 14 trials report on the analyses of the ITT and PP samples clearly, and frequently only one analysis is presented. When this occurs, it is not always the case that the authors justify reporting on only one analysis despite having two samples. When authors do provide a justification, the reasoning centres on the outcomes of the two analyses being equal or non-different, and as such, reporting on both samples is not required.

## 4 Discussion

In the present review, 122 psychology research papers using equivalence analyses, published between 1999 and 2020 were examined. The findings indicate that the use of equivalence analyses is highly inconsistent both across the literature and within specific subdisciplines. These results indicate that definitions of meaningless vary greatly between researchers—additionally, the descriptions associated with how researchers determine what is a meaningless difference (i.e., justifications for margin specification) are often vague and lack enough information to adequately describe the specification process and describe the meaninglessness within the equivalence margin. Finally, the papers reviewed indicate that data complications are an underreported and potentially under-examined component of statistical analyses. Taken together the findings have implications for drawing appropriate inferences from specific tests and threatens the generalisability of analyses in much of the existing literature.

### 4.1 Selection of the equivalence analysis

The first aim of this review was to quantify the most used tests of equivalence (and non-inferiority). This first finding relevant to this research aim is that a substantial proportion of the literature favoured non-inferiority testing. This may be linked to the popularity of the approach in medical and bioequivalence fields, and thus its popularity in psychology may be jointly linked to the availability of applicable educational material and the popularity of non-inferiority testing in other research fields. Additionally, non-inferiority testing involves only specifying the lower half of the equivalence margin because the primary research question being addressed involves demonstration that the mean of one set of scores *is not worse than* the mean from another set of scores. As such, it may be easier to stipulate the margin and justify its specification when the goal is not just to demonstrate that the scores are equivalent, but also that one mean is not much lower than another mean.

The selected test was inadequately or inappropriately described in only six papers of the 122 reviewed. Of these six, only two papers explicitly used superiority testing to examine equivalence of scores (Blom et al. 2015; de Zwaan et al. 2012). Finding evidence that only two papers from this review incorrectly applied superiority testing to issues of equivalency is an indication of increased statistical sophistication within the field and is a largely positive finding, despite issues.

### 4.2 Equivalence margin specification

To address the second aim of this review, the most commonly specified equivalence margins were quantified. The primary finding was that equivalence margin specification was inconsistent and varied widely across papers. The largest issue with this finding is that these inconsistencies can be observed across the literature and within specific sub-disciplines, suggesting little agreement exists regarding the exact nature of meaninglessness. Across the literature, little to no discernible pattern that emerged from examining the equivalence margins reported in the studies reviewed. Despite comparing across several studies, this present review found no evidence of a set of psychology-specific norms with respects to establishing the equivalence margin. Incongruently, Lange and Freitag (2005) found that roughly one-third of the articles examined used approximately half a standard deviation, although authors expressed their margins in various units. No such trend emerged from the findings in this review. It is possible that, given the relative recency of equivalence analyses in psychology the field requires more time (and more discussion) regarding what constitutes meaninglessness.

Additionally, inconsistencies are also present from more specific perspectives: Driessen et al. (2017) and Ly et al. (2015) provided clear examples of this. Despite examining the same phenomena and using the same outcome measure(s), the equivalence margins specified were not the same. This is not an issue unique to this instance and several examples exist in the present review. This finding suggests that researchers do not agree on what represents meaninglessness with respects to equivalence analyses within specific sub-disciplines. These inconsistencies may indicate that researchers need to afford more deliberation to the equivalence margin. Inconsistency in margin specification is not necessarily problematic, particularly considering that the margin may be expressed in various units (e.g., raw scale units or proportions of mean differences). However, it is imperative that margin specification is clearly explained and justified so that readers can assess its appropriateness. More collaboration within specific research areas would possibly enable a consensus to be reached with respects to an agreed upon meaningless difference. Additionally, it may benefit the field of psychology research if research were to be conducted comparing the relative appropriateness of fixed equivalence margins, which is particularly relevant given more recent educational material suggesting fixed-margins may be appropriate if no existing discussion of meaninglessness is available (Lakens 2017; Meyners 2012).

Currently no formal empirical investigation exists that compares pre-specified equivalence margins such as fixed proportions (e.g., 20% of the reference groups' mean) or a fixed effect size (e.g., $d=0.3$, which is equivocal to a SD of 0.3). In areas such as pharmacokinetics the use of pre-specified margins is considered the gold-standard, and due to the nature of the research, stronger assumptions can be made with respects to the true effect of interventions (e.g., drugs). However, measurement of psychological variables involves more measurement error, and consequently, fewer assumptions can be made regarding the direct effect a manipulation (e.g., intervention) has in the change of scores (Anastasi and Urbina 1997). Currently it is unclear if pre-specified margins can be applied to research fields where measurement error is markedly higher than the field in which equivalence analyses was developed. If pre-specified margins were demonstrably appropriate in psychology, many of the current issues researchers face surrounding margin specification would be resolved. Given this, a comparison of pre-specified, fixed equivalence margins may demonstrate that a such pre-specified margins are appropriate for a wide range of research areas.

### 4.3 Reported justifications for equivalence margin specification

To achieve the third aim of this scoping review, the extent to which researchers provide justification for their equivalence margins was investigated. The goal of the justifications should be to demonstrate that the differences between the groups, within the equivalence margin, are meaningless (Lakens et al. 2018; Lange and Freitag 2005). This present review demonstrated that the justifications for the equivalence margins vary widely and frequently lack necessary information. Several studies failed to provide any justification at all, with just a statement of the equivalence margin. A large proportion of studies provided a description of only one to three resources accessed when considering an appropriate margin. These justifications, however, varied greatly in how much detail was afforded to each resource. In some papers, a list of resources is presented, in other papers, each resource is afforded some elaboration as to why it was selected and how it was useful in establishing the margin. The spectrum of information provided is alarming and suggests that no clear benchmarks exist for researchers to refer to when reporting their analyses. Unfortunately, we considered that only a handful of papers examined provided a sufficiently elaborative approach to justifying the equivalence margin. Papers that provided such a justification are examples of the gold-standard approach, where each resource accessed is listed and described, and a corresponding argument is given as to how the margin reflects a meaningless difference.

In some instances, it is true that resources with which to base the equivalence margin upon are non-existent. For researchers implementing equivalence analyses, particularly outside of clinical psychological research, this may be a frequent issue. Indeed, multiple resources are not a prerequisite of an appropriate margin, and an equivalence margin can be specified in the face of zero available resources. However, researchers should provide an elaborate and transparent justification for their margin and highlight potentially limiting aspects of the margin specification (e.g., a lack of equivalence analyses in a specific subfield, where the true nature of meaninglessness has not been discussed). Linde et al. (2021) recommended that, if discussions of meaningless are absent from a specific field, a Bayesian approach to equivalence analyses may provide a better estimate of equivalence over the frequentist approaches discussed in this present scoping review.

Given that the specification of the margin is arbitrary and non-trivial, the information accessed should be explicated as clearly as possible (Lange and Freitag 2005). In cases where information is vague or missing it is very difficult for the reader to judge the appropriateness of the equivalence test (Cribbie et al. 2004; Lange and Freitag 2005). This poses a risk to the use of the inferences drawn from such tests. This present review recommends that papers afford very careful deliberation and explicit reporting to their justifications. The justification should demonstrate that the researchers considered carefully (1) the available resources, if any and (2) what true meaninglessness would reflect on the outcome measure under investigation.

While it may appear that much of the onus is being placed on the authors of papers to meet these criteria, it is equally the responsibility of journals and reviewers to identify poorly conducted equivalence analyses and ineffectively justified margins. The peer-review process is designed to identify aspects of research that need refinement, and this area should be focused upon in the review process. Papers that do not provide a justification for the equivalence margin in sufficient detail should be queried further, because the inferences are not readily useful or helpful in better understanding phenomena.

### 4.4 Assumption checking and outlier management

The fourth aim of this present review was addressed by quantifying the extent to which researchers report on the assumptions of their selected tests and outlier management processes. Alarmingly, a gross underreporting of information was found concerning such aspects of analyses. A small proportion of studies reviewed made comments about, for example, the shape of distributions (e.g., normality); however these comments are frequently not tied directly to the equivalence analysis conducted. That is, these comments are often discussed in a broad sense, in instances where multiple modelling techniques are used. Due to the broad nature of these comments, it is unclear whether the assumption checking and outlier management was applied to all analyses in a paper, or only select aspects of the analytical plan.

One possible explanation for this underreporting is simply that psychological researchers are unfamiliar with data-related issues that may compromise their selected equivalence analysis. (Hoekstra et al. 2012) demonstrated that psychological researchers are unfamiliar with relevant assumption checking and outlier management processes for typical superiority tests (e.g., t-tests, linear regression). This is surprising given that these common superiority tests are addressed repeatedly in research methods courses in typical psychology undergraduate degrees (Hoekstra et al. 2012). A lack of familiarity with typical tests may suggest that researchers are even less knowledgeable on more novel, less educationally integrated, statistical testing such as those required for equivalence analyses. Contrastingly, researchers are possibly familiar with the assumption checking and outlier management processes—but believe that the modelling techniques are robust to such issues.

In the case of assumptions, research has demonstrated the robustness of certain superiority techniques with respects to violations (Bathke 2004; Bradley 1980; Kohr and Games 1974). Given that aspects of the standard arsenal of statistical techniques psychological researchers use is believed to be robust to non-severe assumption violations, the same logic may have been (incorrectly) extended to equivalence analyses. This is to say that researchers assume that equivalence analyses are robust to assumption violations, or that assumption violations bare little on the validity of the inferences of such tests, and therefore do not check or report on such issues. Additionally, little empirical research exists on the relative performance of various equivalence analyses under the effect of various data-related issues (e.g., severe non-normality; Counsell and Cribbie 2015; Kong et al. 2004; Mangardich and Cribbie 2014; Rusticus and Lovato 2014). Researchers may be motivated to address the assumptions of their associated test but be unfamiliar as to which issues to address or how to address them given the limited educational and simulation research on these topics. As a consequence, researchers may incorrectly assume that their selected equivalence analysis is robust to all data-issues researchers frequently face, and therefore ignore checking their data and reporting on their findings. Finally, it may be that researchers are familiar with the associated assumptions underlying their selected analysis and are aware that such tests are susceptible to increased error rates when subject to violations of such assumptions but checking and management processes are omitted from published manuscripts. This may be because they are deemed unimportant, or at least, less important than other aspects of the manuscript. If, for example, authors are held to stringent word or page count limits by their selected journal(s), they may be motivated to omit assumption checking and outlier management information to meet these standards. If this is true, researchers would be (1) aware of the assumptions of their associated tests and (2) conducting correct and appropriate checks where relevant and choosing to exclude such information for pragmatic reasons. Future research may attempt to address these speculations via qualitative interviews where researchers are tasked with conducting and reporting equivalence analyses. Such an activity may provide insight into aspects of equivalence analyses that are considered important to understanding the outcomes of such tests.

### 4.5 Selecting the analysis sample

This present review achieved its final aim by quantifying the extent and nature of sample selection in clinical trials. As previously discussed, analyses of ITT samples tend to diminish treatment effects. As a result, analyses on the ITT samples may bias researchers toward inferring equivalence of sets of scores incorrectly, and increase the risk of committing a Type I error (i.e., finding equivalence of samples when the effect in the population is non-equivalent; D'Agostino et al. 2003). Given this, it is surprising to find in this present scoping review that most clinical trials report on the ITT sample alone. These reporting habits appear consistent with recommendations for reporting clinical trials for superiority designs, where ITT is believed to safeguard against Type I errors (Moher et al. 2012). However, these reporting habits are inconsistent with the literature specially related to equivalence designs, where the opposite is true (Gupta 2011; Piaggio et al. 2006, 2012). Given the relative novelty of equivalence designs in psychological literature, researchers may be unaware of the key differences in sample selection for such designs, and may default to familiar sample selection processes, thus using only the ITT sample for analysis.

In the context of existing recommendations for equivalence designs from medical literature, this present scoping review endorses reporting analyses of both the ITT and PP samples (Gøtzsche 2006; Le Henanff et al. 2006; Piaggio et al. 2012). In doing so, authors can demonstrate differences in the efficacy of their treatment arms. The ITT sample analysis should provide an estimate of the overall treatment effect. This treatment effect is likely to be consistent with what can be expected for individuals engaged with a treatment, because barriers to perfect treatment adherence often exist that will manifest with the ITT sample (Kay 2014; Matsuyama 2010). Contrastingly, the PP sample analysis should provide an estimate of the true treatment effect. This true treatment effect is likely to reflect "perfect" treatment adherence, and thus provide slightly different information than what is offered by only reporting the ITT sample analysis (D'Agostino et al. 2003; Kay 2014).

### 4.6 Limitations

While the search string used to gather appropriate research articles in this present review was sufficiently comprehensive such that saturation of information was likely attained, it is possible that a subsection of the literature was missed. This subsection holds papers that conducted equivalence analyses but did not include this information in the title, abstract or keywords of their manuscript. The proportion of the literature that would engage in this type of reporting is difficult to estimate. However, it can be argued that this present review was not concerned with papers of this nature, because its primary focus was on research where a core component of the statistical protocol involved an equivalence analysis. For this reason, it is reasonable to believe that saturation of information was achieved, and the potential subsection of research articles missed by this search string would have added little to the overall findings presented here. To address this limitation, the search string could be combined with articles that have cited popular educational material related to equivalence testing (e.g., Lakens 2017; Quertemont 2011; Rogers et al. 1993; Schuirmann 1987).

### 4.7 Future directions

The findings from this review indicate several directions future research could take. From the broadest perspective, the reporting habits of researchers is very inconsistent and frequently vague or missing information. To gain insight into the cause of these inconsistencies, a qualitative investigation of researchers' use of equivalence analyses could be conducted. This could provide insight into decision cues that lead researchers to conduct such analyses, and the decision-making processes researchers undertake as they navigate through their analyses (e.g., establish the equivalence margin, assess their data and interpret the tests). This may provide insight into the aspects of equivalence analyses researchers struggle with or lack a clear understanding of. Qualitative research with a focus on a bottom-up approach to understanding these processes may enable the development of specific, targeted educational material that most readily addresses the issues psychological researchers face when conducting equivalence analyses.

Another area for future research concerns the equivalence margin. This present study found evidence that equivalence margins vary widely: This is true for research conducted within specific sub-disciplines, as well as across sub-disciplines. In non-psychology research fields the equivalence margin is heavily influenced by external factors such as the

FDA (FDA 1992). Consequently, research in areas such as pharmacokinetics frequently use the same margin (e.g., 20% of the reference group mean). However, it is unclear if this approach could be readily and appropriately adopted in psychology. If the application of a pre-specified margin could be shown to be appropriate in psychology issues surrounding margin inconsistencies would be largely resolved. Given this, there is room to select several pre-specified margins (e.g., 20% of the reference group mean, half a standard deviation of the reference group (e.g., SD = 0.5) or a prespecified effect size (e.g., $d = 0.3$) and compare the feasibility of these approaches. This is a particularly timely point because more recent educational materials have suggested the use of pre-specified margins if the specific research area is lacking existing discussion on meaningless differences that would be used to inform equivalence margin specification (Lakens 2017).

An additional area for future research focuses upon assumption testing and the effects of various data related issues (e.g., sample size). Limited research exists with respect to assumption checking and handling for equivalence analyses. The existing literature addresses several different issues, however lacks a cohesive base from which clear guidelines can be generated (Linde et al. 2021; Mangardich and Cribbie 2014; Rusticus and Lovato 2014; van Wieringen and Cribbie 2014). As such, one direction for future research involves investigation of the effects of various data-related issues (e.g., non-normality, sample size, heterogenous group variance) on the Type I and II error rates of various equivalence analyses. From a practical perspective, combining the investigation of these factors with the testing of pre-specified margins (as mentioned above) would provide a cohesive body of simulation research with which to offer some clear guidelines for conducting equivalence testing for psychology researchers. In conducting simulation research addressing these issues cohesively, and in conjunction with existing literature, a set of cohesive, clear guidelines for assumption checking and management protocols could be derived. An empirical undertaking of this nature would aid to increase statistical sophistication of psychological research, and researchers would be better able to apply equivalence analyses correctly.

## 5 Conclusion

The extant psychological literature using equivalence analyses appears highly inconsistent. The aim of this scoping review was to provide a cohesive mapping of the literature to identify areas that researchers struggle with when conducting equivalence analyses. To this end, several critical aspects of equivalence analyses were shown to be poorly reported, vague or information poor, and several areas for future research were identified consequently. One area for future research identified involves a qualitative approach to understanding researchers' use of equivalence analyses. Another involves simulation research to address existing gaps within the literature regarding the performance of specific equivalence (and non-inferiority) tests under various circumstances (e.g., the presence or absence of various assumption violations). This research has far-reaching implications and applications within the meta-scientific community, where the primary goals are to identify areas for improvement in statistical inference for psychological researchers and provide guidance to psychological researchers. While the use of equivalence analyses may initially appear simple (particularly when using tests such as the TOST procedure), several critical judgements must be made that, if performed incorrectly, may jeopardise the validity of the analysis. The findings from this review indicate the likelihood of misapplication or misinterpretation of findings appear high given the issues we found in the existing literature.

**Author contributions** All authors contributed to the study conception and design. Material preparation, data collection, analysis and interpretation were performed by AM. The first draft of the manuscript was written by AM and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Data availability** All charted data can be found on the Open Science Framework repository https://osf.io/v7j5z/?view_only=8de42816bea64215b4500392f2de8922

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

## References

Acierno, R., Knapp, R., Tuerk, P., Gilmore, A.K., Lejuez, C., Ruggiero, K., Muzzy, W., Egede, L., Hernandez-Tejada, M.A., Foa, E.B.: A non-inferiority trial of prolonged exposure for posttraumatic stress disorder: in person versus home-based telehealth. Behav. Res. Therapy (2017). https://doi.org/10.1016/j.brat.2016.11.009

Alavijeh, M.S., Chishty, M., Qaiser, M.Z., Palmer, A.M.: Drug metabolism and pharmacokinetics, the blood-brain barrier, and central nervous system drug discovery. NeuroRx **2**(4), 554–571 (2005). https://doi.org/10.1602/neurorx.2.4.554

Alfano, C.: Are children with "pure" generalised anxiety disorder impaired? A comparison with comorbid and healthy children. J. Clin. Child Adolesc. Psychol. **41**(6), 739–745 (2012). https://doi.org/10.1080/15374416.2012.715367

Althunian, T.A., de Boer, A., Klungel, O.H., Insani, W.N., Groenwold, R.H.: Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review. Trials **18**(1), 1–9 (2017). https://doi.org/10.1186/s13063-017-1859-x

Anastasi, A., Urbina, S.: Psychological Testing. Prentice Hall/Pearson Education, New York (1997)

Andersson, G., Hesser, H., Veilord, A., Svedling, L., Andersson, F., Sleman, O., Mauritzson, L., Sarkohi, A., Claesson, E., Zetterqvist, V., Lamminen, M., Eriksson, T., Carlbring, P.: Randomised controlled non-inferiority trial with 3-year follow-up of internet-delivered versus face-to-face group cognitive behavioural therapy for depression. J. Affect. Disord. **151**(3), 986–994 (2013). https://doi.org/10.1016/j.jad.2013.08.022

Arksey, H., O'Malley, L.: Scoping studies: towards a methodological framework. Int. J. Soc. Res. Methodol. **8**(1), 19–32 (2005). https://doi.org/10.1080/1364557032000119616

Bakker, M., Wicherts, J.M.: Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. Psychol. Methods **19**(3), 409 (2014). https://doi.org/10.1037/met0000014

Ball, L.C., Cribbie, R.A., Steele, J.R.: Beyond gender differences: using tests of equivalence to evaluate gender similarities. Psychol. Women Q. **37**(2), 147–154 (2013). https://doi.org/10.1177/0361684313480483

Barlow, D.H., Farchione, T.J., Bullis, J.R., Gallagher, M.W., Murray-Latin, H., Sauer-Zavala, S., Bentley, K.H., Thompson-Hollands, J., Conklin, L.R., Boswell, J.F., Ametaj, A., Carl, J.R., Boettcher, H.T., Cassiello-Robbins, C.: The unified protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: a randomized clinical trial. JAMA Psychiat. **74**(9), 875–884 (2017). https://doi.org/10.1001/jamapsychiatry.2017.2164

Bathke, A.: The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data. J. Stat. Plan. Inference **126**(2), 413–422 (2004). https://doi.org/10.1016/j.jspi.2003.09.010

Bauer, B.W., Gai, A.R., Duffy, M.E., Rogers, M.L., Khazem, L.R., Martin, R.L., Joiner, T.E., Capron, D.W.: Fearlessness about death does not differ by suicide attempt method. J. Psychiatr. Res. **124**, 42–49 (2020). https://doi.org/10.1016/j.jpsychires.2020.02.014

Beck, B.D., Lund, S.T., Søgaard, U., Simonsen, E., Tellier, T.C., Cordtz, T.O., Laier, G.H., Moe, T.: Music therapy versus treatment as usual for refugees diagnosed with posttraumatic stress disorder (PTSD): study protocol for a randomized controlled trial. Trials **19**(1), 301–320 (2018). https://doi.org/10.1186/s13063-018-2662-z

Beukes, E.W., Andersson, G., Allen, P.M., Manchaiah, V., Baguley, D.M.: Effectiveness of guided internet-based cognitive behavioral therapy vs face-to-face clinical care for treatment of tinnitus: a randomized clinical trial. JAMA Otolaryngol. Head Neck Surg. **144**(12), 1126–1133 (2018). https://doi.org/10.1001/jamaoto.2018.2238

Blom, K., Tarkian Tillgren, H., Wiklund, T., Danlycke, E., Forssen, M., Soderstrom, A., Johansson, R., Hesser, H., Jernelov, S., Lindefors, N., Andersson, G., Kaldo, V.: Internet- vs. group-delivered cognitive behavior therapy for insomnia: a randomized controlled non-inferiority trial. Behav. Res. Ther. **70**, 47–55 (2015). https://doi.org/10.1016/j.brat.2015.05.002

Blumberger, D.M., Vila-Rodriguez, F., Thorpe, K.E., Feffer, K., Noda, Y., Giacobbe, P., Knyahnytska, Y., Kennedy, S.H., Lam, R.W., Daskalakis, Z.J., Downar, J.: Effectiveness of theta burst versus high-frequency repetitive transcranial magnetic stimulation in patients with depression (THREE-D): a randomised non-inferiority trial: Erratum. The Lancet **391**(10139), 1683–1692 (2018). https://doi.org/10.1016/S0140-6736(18)30295-2

Bradley, J.V.: Nonrobustness in Z, t, and F tests at large sample sizes. Bull. Psychon. Soc. **16**(5), 333–336 (1980)

Bramoweth, A.D., Lederer, L.G., Youk, A.O., Germain, A., Chinman, M.J.: Brief behavioral treatment for insomnia vs. cognitive behavioral therapy for insomnia: results of a randomized noninferiority clinical trial among veterans. Behav. Ther. **51**(4), 535–547 (2020). https://doi.org/10.1016/j.beth.2020.02.002

Charig, R., Moghaddam, N.G., Dawson, D.L., Merdian, H.L., das Nair, R.: A lack of association between online pornography exposure, sexual functioning, and mental well-being. Sex. Relatsh. Ther. **35**(2), 258–281 (2020). https://doi.org/10.1080/14681994.2020.1727874

Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum Associates, Mahwah (1988)

Counsell, A., Cribbie, R.A.: Equivalence tests for comparing correlation and regression coefficients. Br. J. Math. Stat. Psychol. **68**(2), 292–309 (2015). https://doi.org/10.1111/bmsp.12045

Cribbie, R.A., Gruman, J.A., Arpin-Cribbie, C.A.: Recommendations for applying tests of equivalence. J. Clin. Psychol. **4**(4), 1–10 (2004). https://doi.org/10.1002/jclp.10217

D'Agostino, R.B., Massaro, J.M., Sullivan, L.M.: Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. Stat. Med. **22**(2), 169–186 (2003). https://doi.org/10.1002/sim.1425

de Zwaan, M., Herpertz, S., Zipfel, S., Tuschen-Caffier, B., Friederich, H.C., Schmidt, F., Gefeller, O., Mayr, A., Lam, T., Schade-Brittinger, C., Hilbert, A.: INTERBED: internet-based guided self-help for overweight and obese patients with full or subsyndromal binge eating disorder. A multicenter randomized controlled trial. Trials **13**(1), 220 (2012). https://doi.org/10.1186/1745-6215-13-220

Dirkse, D., Hadjistavropoulos, H.D., Alberts, N.A., Karin, E., Schneider, L.H., Titov, N., Dear, B.F.: Making Internet-delivered cognitive behaviour therapy scalable for cancer survivors: a randomized non-inferiority trial of self-guided and technician-guided therapy. J. Cancer Surviv. Res. Pract. **14**(2), 211–225 (2020). https://doi.org/10.1007/s11764-019-00810-9

Driessen, E., Van, H.L., Peen, J., Don, F.J., Twisk, J.W.R., Cuijpers, P., Dekker, J.J.M.: Cognitive-behavioral versus psychodynamic therapy for major depression: secondary outcomes of a randomized clinical trial. J. Consult. Clin. Psychol. **85**(7), 653–663 (2017). https://doi.org/10.1037/ccp0000207

Dunn, M.E., Fried-Somerstein, A., Flori, J.N., Hall, T.V., Dvorak, R.D.: Reducing alcohol use in mandated college students: a comparison of a brief motivational intervention (BMI) and the expectancy challenge alcohol literacy curriculum (ECALC). Exp. Clin. Psychopharmacol. **28**(1), 87–98 (2019). https://doi.org/10.1037/pha0000290

Eschenbeck, H., Lehner, L., Hofmann, H., Bauer, S., Becker, K., Diestelkamp, S., Kaess, M., Moessner, M., Rummel-Kluge, C., Salize, H.J., Thomasius, R., Bertsch, K., Bilic, S., Brunner, R., Feldhege, J., Gallinat, C., Herpertz, S.C., Koenig, J., Lustig, S., et al.: School-based mental health promotion in children and adolescents with StresSOS using online or face-to-face interventions: Study protocol for a randomized controlled trial within the ProHEAD Consortium. Trials **20**(1), 12–64 (2019). https://doi.org/10.1186/s13063-018-3159-5

Fals-Stewart, W., Klostermann, K., O'Farrell, T.J., Yates, B.T., Birchler, G.R.: Brief relationship therapy for alcoholism: a randomized clinical trial examining clinical efficacy and cost-effectiveness. Psychol. Addict. Behav. **19**(4), 363–371 (2005). https://doi.org/10.1037/0893-164X.19.4.363

Fals-Stewart, W., Lam, W.K.K.: Brief behavioral couples therapy for drug abuse: a randomized clinical trial examining clinical efficacy and cost-effectiveness. Fam. Syst. Health **26**(4), 377–392 (2008). https://doi.org/10.1037/1091-7527.26.4.377

FDA. (1992). Points to consider: clinical development and labeling of anti-infective drug products. U.S. Dep. of Health and Human Services.

Forand, N.R., Feinberg, J.E., Barnett, J.G., Strunk, D.R.: Guided internet CBT versus "gold standard" depression treatments: an individual patient analysis. J. Clin. Psychol. **75**(4), 581–593 (2019). https://doi.org/10.1002/jclp.22733

Gagnon, S., Marshall, S., Kadulina, Y., Stinchcombe, A., Bedard, M., Gelinas, I., Man-Son-Hing, M., Mazer, B., Naglie, G., Porter, M.M., Rapoport, M., Tuokko, H., Vrkljan, B., Candrive Research T: CIHR candrive cohort comparison with Canadian household population holding valid driver's licenses. Can. J. Aging **35**(1), 99–109 (2016). https://doi.org/10.1017/S0714980816000052

Glass, G.V., Peckham, P.D., Sanders, J.R.: Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev. Educ. Res. **42**(3), 237–288 (1972). https://doi.org/10.3102/00346543042003237

Goldstein, L.A., Adler Mandel, A.D., DeRubeis, R.J., Strunk, D.R.: Outcomes, skill acquisition, and the alliance: similarities and differences between clinical trial and student therapists. Behav. Res. Ther. **129**, 103608 (2020). https://doi.org/10.1016/j.brat.2020.103608

Goodman, J.A., Israel, T.: An online intervention to promote predictors of supportive parenting for sexual minority youth. J. Fam. Psychol. **34**(1), 90–100 (2020). https://doi.org/10.1037/fam0000614

Gøtzsche, P.C.: Lessons from and cautions about noninferiority and equivalence randomized trials. JAMA **295**(10), 1172–1174 (2006). https://doi.org/10.1001/jama.295.10.1172

Grose Deters, F., Mehl, M.R., Eid, M.: Narcissistic power poster? On the relationship between narcissism and status updating activity on Facebook. J. Res. Person. **53**(165), 174 (2014). https://doi.org/10.1016/j.jrp.2014.10.004

Gross, D., Belcher, H.M.E., Budhathoki, C., Ofonedu, M.E., Dutrow, D., Uveges, M.K., Slade, E.: Reducing preschool behavior problems in an urban mental health clinic: a pragmatic, non-inferiority trial. J. Res. Pers. **58**(6), 572–581 (2019). https://doi.org/10.1016/j.jaac.2018.08.013

Gupta, S.K.: Intention-to-treat concept: a review. Perspect. Clin. Res. **2**(3), 109 (2011). https://doi.org/10.4103/2229-3485.83221

Hedman, E., Andersson, G., Ljotsson, B., Andersson, E., Ruck, C., Mortberg, E., Lindefors, N.: Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: a randomized controlled non-inferiority trial. PLoS ONE **6**(3), e18001 (2011). https://doi.org/10.1371/journal.pone.0018001

Hedman, E., El Alaoui, S., Lindefors, N., Andersson, E., Rück, C., Ghaderi, A., Kaldo, V., Lekander, M., Andersson, G., Ljótsson, B.: Clinical effectiveness and cost-effectiveness of internet- vs. group-based cognitive behavior therapy for social anxiety disorder: 4-year follow-up of a randomized trial. Behav. Res. Ther. **59**, 20–29 (2014). https://doi.org/10.1016/j.brat.2014.05.010

Herbert, M.S., Afari, N., Liu, L., Heppner, P., Rutledge, T., Williams, K., Eraly, S., VanBuskirk, K., Nguyen, C., Bondi, M., Atkinson, J.H., Golshan, S., Wetherell, J.L.: Telehealth versus in-person acceptance and commitment therapy for chronic pain: a randomized noninferiority trial. J. Pain **18**(2), 200–211 (2017). https://doi.org/10.1016/j.jpain.2016.10.014

Hoekstra, R., Kiers, H., Johnson, A.: Are assumptions of well-known statistical techniques checked, and why (not)? Front. Psychol. **3**, 137 (2012). https://doi.org/10.3389/fpsyg.2012.00137

Hofmann, S.G., Curtiss, J., Khalsa, S.B.S., Hoge, E., Rosenfield, D., Bui, E., Keshaviah, A., Simon, N.: Yoga for generalized anxiety disorder: design of a randomized controlled clinical trial. Contemp. Clin. Trials **44**, 70–76 (2015). https://doi.org/10.1016/j.cct.2015.08.003

Jongen, S., Vermeeren, A., van der Sluiszen, N.N.J.J.M.D., Schumacher, M.B., Theunissen, E.L., Kuypers, K.P.C., Vuurman, E.F.P.M., Ramaekers, J.G.: A pooled analysis of on-the-road highway driving studies in actual traffic measuring standard deviation of lateral position (i.e., "weaving") while driving at a blood alcohol concentration of 0.5 g/L. Psychopharmacology **234**(5), 837–844 (2017). https://doi.org/10.1007/s00213-016-4519-z

Kay, R.: Statistical Thinking for Non-statisticians in Drug Regulation. Wiley, Hoboken (2014)

Kohr, R.L., Games, P.A.: Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. J. Exp. Educ. **43**(1), 61–69 (1974). https://doi.org/10.1080/00220973.1974.10806305

Kong, L., Kohberger, R.C., Koch, G.G.: Type I error and power in noninferiority/equivalence trials with correlated multiple endpoints: an example from vaccine development trials. J. Biopharm. Stat. **14**(4), 893–907 (2004). https://doi.org/10.1081/BIP-200035454

Kuang, J., Milhorn, H., Stuppy-Sullivan, A., Jung, S., Yi, R.: Alternate versions of a fixed-choice, delay-discounting assessment for repeated-measures designs. Exp. Clin. Psychopharmacol. **26**(5), 503–508 (2018). https://doi.org/10.1037/pha0000211

Lakens, D.: Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Person. Sci. **8**(4), 355–362 (2017). https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A.M., Isager, P.M.: Equivalence testing for psychological research: a tutorial. Adv. Methods Pract. Psychol. Sci. **1**(2), 259–269 (2018). https://doi.org/10.1177/2515245918770963

Lange, S., Freitag, G.: Choice of delta: requirements and reality-results of a systematic review. Biom. J. **47**(1), 12–27 (2005). https://doi.org/10.1002/bimj.200410085

Le Henanff, A., Giraudeau, B., Baron, G., Ravaud, P.: Quality of reporting of noninferiority and equivalence randomized trials. JAMA **295**(10), 1147–1151 (2006)

Leichsenring, F., Abbass, A., Driessen, E., Hilsenroth, M., Luyten, P., Rabung, S., Steinert, C.: Equivalence and non-inferiority testing in psychotherapy research. Psychol. Med. **48**(11), 1917–1919 (2018). https://doi.org/10.1017/S0033291718001289

Linde, M., Tendeiro, J.N., Selker, R., Wagenmakers, E.-J., van Ravenzwaaij, D.: Decisions about equivalence: a comparison of TOST, HDI-ROPE, and the Bayes factor. Psychol. Methods (2021). https://doi.org/10.1037/met0000402

Liu, L., Thorp, S.R., Moreno, L., Wells, S.Y., Glassman, L.H., Busch, A.C., Zamora, T., Rodgers, C.S., Allard, C.B., Morland, L.A., Agha, Z.: Videoconferencing psychotherapy for veterans with PTSD: results from a randomized controlled non-inferiority trial. J. Telemed. Telecare (2019). https://doi.org/10.1177/1357633X19853947

Ly, K.H., Topooco, N., Cederlund, H., Wallin, A., Bergstrom, J., Molander, O., Carlbring, P., Andersson, G.: Smartphone-supported versus full behavioural activation for depression: a randomised controlled trial. PLoS ONE (2015). https://doi.org/10.1371/journal.pone.0126559

Malinvaud, D., Londero, A., Niarra, R., Peignard, P., Warusfel, O., Viaud-Delmon, I., Chatellier, G., Bonfils, P.: Auditory and visual 3D virtual reality therapy as a new treatment for chronic subjective tinnitus: results of a randomized controlled trial. Hear. Res. **333**, 127–135 (2016). https://doi.org/10.1016/j.heares.2015.12.023

Mangardich, H., Cribbie, R.A.: Assessing clinical significance using robust normative comparisons. Psychother. Res. **25**(2), 239–248 (2014). https://doi.org/10.1080/10503307.2014.889329

Mathersul, D.C., Tang, J.S., Jay Schulz-Heik, R., Avery, T.J., Seppälä, E.M., Bayley, P.J.: Study protocol for a non-inferiority randomised controlled trial of SKY breathing meditation versus cognitive processing therapy for PTSD among veterans. BMJ Open (2019). https://doi.org/10.1136/bmjopen-2018-027150

Mathiasen, K., Andersen, T.E., Riper, H., Kleiboer, A.A., Roessler, K.K.: Blended CBT versus face-to-face CBT: a randomised non-inferiority trial. BMC Psychiatry **16**(1), 432 (2016). https://doi.org/10.1186/s12888-016-1140-y

Matsuyama, Y.: A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. Stat. Med. **29**(20), 2107–2116 (2010). https://doi.org/10.1002/sim.3987

Meyners, M.: Equivalence tests: a review. Food Qual. Prefer. **26**(2), 231–245 (2012). https://doi.org/10.1016/j.foodqual.2012.05.003

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P., Elbourne, D., Egger, M., Altman, D.G.: CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. J. Clin. Epidemiol. **63**(8), 1–37 (2012). https://doi.org/10.1016/j.jclinepi.2010.03.004

Murray, L.K., Haroz, E.E., Doty, S.B., Singh, N.S., Bogdanov, S., Bass, J., Dorsey, S., Bolton, P.: Testing the effectiveness and implementation of a brief version of the common elements treatment approach (CETA) in Ukraine: a study protocol for a randomized controlled trial. Trials **19**(1), 418 (2018). https://doi.org/10.1186/s13063-018-2752-y

Nimon, K.F.: Statistical assumptions of substantive analyses across the general linear model: a mini-review. Front. Psychol. **3**, 322 (2012). https://doi.org/10.3389/fpsyg.2012.00322

Norton, P.J., Barrera, T.L.: Transdiagnostic versus diagnosis-specific cbt for anxiety disorders: a preliminary randomized controlled noninferiority trial. Depress. Anxiety **29**(10), 874–882 (2012). https://doi.org/10.1002/da.21974

Piaggio, G., Elbourne, D.R., Altman, D.G., Pocock, S.J., Evans, S.J.W.: Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. J. Am. Med. Assoc. **295**(10), 1152–1160 (2006). https://doi.org/10.1001/jama.295.10.1152

Piaggio, G., Elbourne, D.R., Pocock, S.J., Evans, S.J., Altman, D.G.: Reporting of noninferiority and equivalence: randomized trials extension of the CONSORT 2010 statement. J. Am. Med. Assoc. **308**(24), 2594–2604 (2012). https://doi.org/10.1001/jama.2012.87802

Pong, S., Urner, M., Fowler, R.A., Mitsakakis, N., Seto, W., Hutchison, J.S., Daneman, N.: Testing for non-inferior mortality: a systematic review of non-inferiority margin sizes and trial characteristics. BMJ Open **11**(4), e044480 (2021). https://doi.org/10.1136/bmjopen-2020-044480

Quertemont, E.: How to statistically show the absence of an effect. Psychologica Belgica (2011). https://doi.org/10.5334/pb-51-2-109

Rehal, S., Morris, T.P., Fielding, K., Carpenter, J.R., Phillips, P.P.: Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. BMJ Open (2016). https://doi.org/10.1136/bmjopen-2016-012594

Rogers, J.L., Howard, K.I., Vessey, J.T.: Using significance tests to evaluate equivalence between two experimental groups. Psychol. Bull. **113**(3), 553–565 (1993). https://doi.org/10.1037/0033-2909.113.3.553

Romijn, G., Riper, H., Kok, R., Donker, T., Goorden, M., van Roijen, L.H., Kooistra, L., van Balkom, A., Koning, J.: Cost-effectiveness of blended vs. face-to-face cognitive behavioural therapy for severe anxiety disorders: study protocol of a randomized controlled trial. BMC Psychiatry **15**(1), 311 (2015). https://doi.org/10.1186/s12888-015-0697-1

Rusticus, S.A., Lovato, C.Y.: Impact of sample size and variability on the power and type I error rates of equivalence tests: a simulation study. Pract. Assess. Res. Eval. **19**(1), 11 (2014). https://doi.org/10.7275/4s9m-4e81

Schmitz, N., Hartkamp, N., Brinschwitz, C., Michalek, S., Tress, W.: Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a randomized trial. Acta Psychiatr. Scand. **102**(2), 147–152 (2001). https://doi.org/10.1034/j.1600-0447.2000.102002147.x

Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharm. Biopharm. **15**(6), 657–680 (1987). https://doi.org/10.1007/BF01068419

Schumi, J., Wittes, J.T.: Through the looking glass: Understanding non-inferiority [review]. Trials (2011). https://doi.org/10.1186/1745-6215-12-106

Sloan, D.M., Marx, B.P., Lee, D.J., Resick, P.A.: A brief exposure-based treatment vs cognitive processing therapy for posttraumatic stress disorder: a randomized noninferiority clinical trial. JAMA Psychiat. **75**(3), 233–239 (2018). https://doi.org/10.1001/jamapsychiatry.2017.4249

Smiti, A.: A critical overview of outlier detection methods. Comput. Sci. Rev. **38**, 100306 (2020). https://doi.org/10.1016/j.cosrev.2020.100306

Srivastava, A.: Effect of non-normality on the power of the analysis of variance test. Biometrika **46**(1/2), 114–122 (1959). https://doi.org/10.2307/2332813

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D., Horsley, T., Weeks, L.: PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann. Intern. Med. **169**(7), 467–473 (2018). https://doi.org/10.7326/M18-0850

van Wieringen, K., Cribbie, R.A.: Evaluating clinical significance: Incorporating robust statistics with normative comparison tests. Br. J. Math. Stat. Psychol. **67**(2), 213–230 (2014). https://doi.org/10.1111/bmsp.12015

Walker, E., Nowacki, A.S.: Understanding equivalence and noninferiority testing. J. Gen. Intern. Med. **26**(2), 192–196 (2011). https://doi.org/10.1007/s11606-010-1513-8

Wangge, G., Klungel, O.H., Roes, K.C., De Boer, A., Hoes, A.W., Knol, M.J.: Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. PLoS ONE **5**(10), e13550 (2010). https://doi.org/10.1371/journal.pone.0013550

Weinstock, J., April, L.M., Kallmi, S.: Is subclinical gambling really subclinical? Addict. Behav. **73**, 185–191 (2017). https://doi.org/10.1016/j.addbeh.2017.05.014

Westlake, W.J.: Use of confidence intervals in analysis of comparative bioavailability trials. J. Pharm. Sci. **61**(8), 1340–1341 (1972). https://doi.org/10.1002/jps.2600610845

Witte, R.S., Witte, J.S.: Statistics, 11th edn. Wiley, Hoboken (2017)

Yeung, T., Martin, J.L., Fung, C.H., Fiorentino, L., Dzierzewski, J.M., Tapia, J.C.R., Song, Y., Josephson, K., Jouldjian, S., Mitchell, M.N., Alessi, C.: Sleep outcomes with cognitive behavioral therapy for insomnia are similar between older adults with low vs. high self-reported physical activity. Front. Aging Neurosci. **10**(1), 274 (2018). https://doi.org/10.3389/fnagi.2018.00274

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.