# Does Answering Survey Questions About Mental Health Cause Distress?

Larry Paing[1], Katie O'Keefe[1], Brandon McMurtrie[12] and Matt N. Williams[1]

[1]Massey University.

[2]To whom correspondence should be addressed. Email: brandonmcmurtriebg@gmail.com

## Submission Statement

Accepted by *Meta-Psychology*. This article will be published following reproducibility checks and copy-editing. The full editorial process of all articles under review at Meta-Psychology can be found following this link:

https://tinyurl.com/mp-submissions

You will find this preprint by searching for the first author's name.

## Abstract

Self-report surveys measuring mental health constructs such as anxiety, depression, and trauma symptoms are regularly used in psychological research. Researchers and ethics committees are often concerned about the possibility that completing surveys on mental health might itself provoke negative emotional reactions. While prior studies have not found strong evidence of such impacts, the existing literature is limited by a lack of preregistered studies, and a lack of studies applying statistical methods that can establish the *absence* of an effect. In Study 1, we therefore conducted a preregistered within-subjects experiment ($N$ = 337) testing the effects of completing measures of anxiety and depression on state distress, tiredness, happiness, and anxiety. In Study 2, we conducted a between-subjects experiment ($N$ = 871) testing the effects of answering questions about trauma symptoms on state distress, perceived meaningfulness of participation, and unwanted thoughts. In both studies, equivalence tests indicated that the effects of the manipulations on state distress were significantly within our preregistered equivalence bounds of $d \pm 0.2$, albeit that in Study 1 the estimated effect ($d$ = 0.10) was significantly greater than zero. Our findings suggest that answering surveys about mental health topics may provoke some distress. However, these effects appear to be very small, even with outcomes measured immediately after an experimental manipulation.

# Introduction

Self-report surveys are a staple in psychological research and practice, and indeed in many other academic disciplines and areas of commerce. While we know of no precise estimate of the quantity of time humans spend completing surveys each year, no doubt the figure is in the millions of hours. Thus, the information collected forms a crucial part of the evidence base in many disciplines. While self-report surveys certainly have their limitations (Paulhus & Vazire, 2009), they are an indispensable source of evidence about humans, especially in relation to internal experiences such as thoughts, attitudes and emotions. But how does completing a survey affect the survey-taker?

The possibility of *iatrogenic* impacts of completing surveys is especially salient in relation to surveys containing questions about mental health and other sensitive topics. (The term *iatrogenic effects* or *iatrogenesis* refers to the inadvertent adverse impacts of an intervention or an assessment.) The possibility exists that asking participants to consider mental health symptoms could elicit iatrogenic effects such as unintended distress or even exacerbate existing symptoms. Concerns around iatrogenesis have sometimes created significant hurdles for researchers seeking ethical approval to conduct studies on sensitive topics, with institutional review boards (IRBs) and ethics committees being wary of the potential harm that research may cause. IRBs often consider research on sensitive topics like sex, trauma, self-harm, and suicide to be high-risk and needing more precautions than other types of psychological research (see Lakeman & FitzGerald, 2009; Rinehart et al., 2017). As a result, many researchers modify various aspects of their research protocols, including making surveys less detailed (Jaffe et al., 2015). While it is true that ethical research must minimise or justify inherent risks, not asking about mental distress can be costly to scholarship progress. Despite this, some scholars argue that review committees tend to make subjective decisions about risks and benefits without robust empirical data regarding the iatrogenic impacts of sensitive research (Rinehart et al., 2017).

## The Nocebo Effect

One area of theorising that provokes specific reason for concern about the iatrogenic effects of completing mental health surveys is that concerned with the *nocebo* effect. The term nocebo was originally coined in the medical field to delineate the adverse effects of placebo from its positive impact (Hahn, 1997). Placebo and nocebo effects constitute a treatment's nonspecific side effects, which are physiological symptoms with no explicit biological mechanism (Barsky et al., 2002).

In recent years, researchers have applied symptom perception theory (Pennebaker, 1982) to explain nocebo effects. The theory proposes that health beliefs or schemas can influence symptom reporting by enabling selective attention to sensations and making people more introspective towards their internal experiences. It is possible that asking participants to focus on distressing symptoms triggers or increases awareness of such experiences, even if temporarily, potentially in turn causing distress. Furthermore, asking participants about symptoms of mental ill-health might cause them to focus on and ruminate about such symptoms, potentially intensifying their negative emotions (see Nolen-Hoeksema et al., 2008).

## General Empirical Evidence of Iatrogenic Impacts

Despite plausible reasons to suspect that iatrogenic impacts might occur, it is currently less clear whether surveys about mental health topics do in fact have iatrogenic impacts. A scoping review by McMurtrie (2023) provides an evaluation of existing literature on the topic. The review included 37 studies (10 experimental and 27 observational) that employed remote survey methods like written and online administrations. The studies largely examined negative emotional reactions as the dependent variable, using self-report measures that contained single or multiple items. They generally refer to this outcome as distress; however, various constructs or terms are present in the literature, including anxiety, stress, depression, suicidality, discomfort, upset, and regret. Overall, the evidence for iatrogenic impacts was limited, with effects typically being small or non-significant. A minority of participants did report higher distress or a similar affective response in some studies, but these participants usually also expressed a willingness to participate in further research, or that their research participation was meaningful.

## Constructs of Interest for This Study

It is reasonable to suppose that whatever iatrogenic impacts completing a survey may have, these impacts will differ depending on the nature of the questions in the survey. Despite credible evidence that the impacts of suicide screening on distress are limited, iatrogenic impacts of other types of measures are possible. Indeed, iatrogenic impacts have been assessed for surveys across a range of different content domains. There has been a particularly strong focus in the existing literature on the potential impacts of screening for suicide symptoms; meta-analyses of this literature (Blades et al., 2018; DeCou & Schumann, 2018) found no evidence of significant detrimental impacts. However, a great deal of psychological research uses measures of mental health constructs other than suicidality. Three content domains of particular interest in the current study are depression, anxiety, and trauma symptoms. We focus on these domains for several reasons.

First, anxiety, depression, and trauma symptoms are all the subject of a very substantial quantity of ongoing research. For example, a Scopus search on 8 September 2023 produced more than 57,000 documents with the word “depression” in the title, abstract or keywords published in 2022 alone. A similar search for “anxiety” produces more than 46,000 documents, and trauma more than 25,000. The volume of research conducted on these topics (a substantial subset of which involves self-report surveys) means that any iatrogenic impacts are a relatively high-stakes concern.

Second, depression, anxiety, and trauma symptoms are all relatively common experiences. Prevalence estimates for 2019 suggest approximately 301 million people living with anxiety and 280 million with depression globally (World Health Organization, 2022). Likewise, many people experience trauma in their lifetime. The World Health Organization’s survey of over 68,000 individuals across 24 countries revealed that 70.4% of respondents reported a trauma experience with the highest risk for post-traumatic stress disorder (PTSD) from intimate partner violence and sexual violence (Kessler et al., 2017). Because these conditions are very prevalent, many participants completing surveys about symptoms of depression, anxiety or trauma may indeed be experiencing such symptoms, increasing the plausibility of negative effects from their attention being drawn to these symptoms.

Third, depression, anxiety, and trauma symptoms are generally negative or harmful phenomena. It is more plausible that a participant might be negatively affected by answering questions about symptoms of depression, their personal traumatic experiences, or objects of great fear, than by surveys with less sensitive content (e.g., pertaining to interests, attitudes towards commercial products, or personality dimensions).

## Evidence for Iatrogenic Impacts of Depression and Anxiety Surveys

There is currently very little experimental research focusing on the iatrogenic impacts of depression and anxiety measures in particular. In an experimental study with Melbourne high school students, Robinson et al. (2011) tested the effect of screening students for suicidal ideation, deliberate self-harm, and distress (including symptoms of anxiety and depression). They found no evidence that screening affects negative emotional states such as anger, depression and confusion[1]. Another experiment with 2,342 high school students by Gould et al.

---

[1] They did find that the screening survey produced a significant increase in “vigor” scores.

(2005) found that asking students to complete measures of depression, substance use problems and suicide attempt history had no significant effect on distress either immediately after the survey or two days later.

Neither of these studies permits the effects of anxiety and depression questions to be distinguished from those of questions about other constructs such as suicidality and substance use. However, the absence of evidence of effects even when the stimuli included questions about these other constructs suggests the effects of answering questions about anxiety and depression are unlikely to be large.

## Evidence for Iatrogenic Impacts of Trauma Surveys

The inclusion of questions about trauma or trauma symptoms can often cause special concerns for institutional review boards (IRBs) or ethics committees (see Jaffe et al., 2015; Newman, 2008; Rinehart et al., 2017; Yeater et al., 2012). There is even the possibility that such surveys might constitute psychological *retraumatisation*, which occurs when individuals with a trauma history are exposed to a situation that triggers the re-experiencing of their trauma (see Courtois, 2012).

There has therefore been more research focusing on the effects of questions about trauma or trauma symptoms than is the case for anxiety and depression. In a study of 127 patients receiving care for PTSD and related conditions, Ferrier-Auerbach et al. (2009) found that participants who completed a trauma survey battery reported being more upset compared to participants who completed a non-trauma survey battery. Although statistically significant, the difference was less than one point on a nine-point scale. Similarly, Yeater et al. (2012) found that their trauma and sex survey batteries produced slightly higher negative emotions than the control group.

In a particularly informative study, Cook et al. (2015) conducted an experiment study with five sensitive-survey conditions, including stressful, stigmatising, traumatic, and sexual items. The study produced few significant effects of sensitive surveys in comparison to a control condition, bar a small and temporary decrease in positive affect for participants answering questions about sexually violating events.

Overall, previous experimental studies have indicated that trauma surveys can cause some increase in distress or a decrease in positive emotion. However, the effect sizes have tended to be small.

## Rationale for the Current Studies

Despite a body of existing work on the iatrogenic impacts of surveys (including those about depression, anxiety, and trauma symptoms), the existing literature has several limitations.

First, none of the existing experimental studies we reviewed reported an *a priori* power analysis. Given that the iatrogenic impacts of surveys seem unlikely to be very large, ensuring adequate statistical power to detect fairly small effects is crucial.

Second, the existing literature has relied heavily on null hypothesis significance testing (NHST) where the null hypothesis is of a *zero* effect. However, a non-significant result in such a test is not evidence of the absence of an effect (Lakens et al., 2018). We know of no existing study on iatrogenic impacts of anxiety, depression, or trauma surveys that has specifically employed an analysis strategy that can provide evidence for the *absence* of an effect. Such strategies include equivalence testing (Lakens et al., 2018), Bayes factors (Kass & Raftery, 1995; Rouder et al., 2009; Williams et al., 2017), and Bayesian estimation with a region of practical equivalence (Kruschke, 2013).

Third, to our knowledge, no *preregistered* study on iatrogenic impacts of depression, anxiety, or trauma surveys has previously been reported. Preregistration is the practice of specifying a research design, hypotheses, data collection and planned analyses in a durable repository before collecting data (or sometimes after data collection but before analysis; Akker et al., 2021). While the merits of preregistration have been subject to meta-scientific debate (see Nosek et al., 2019; Rubin, 2020; Szollosi et al., 2020), preregistration can provide readers with a better capacity to evaluate whether a hypothesis was subjected to a severe test (Lakens, 2019).

Consequently, the current project sought to investigate whether surveys on depression, anxiety (Study 1), and trauma symptoms (Study 2) would cause distress among participants. In both studies, we focused primarily on the relatively generic outcome variable of state **distress**, given that impacts on distress have been a significant focus in the existing literature, but included some supplementary hypotheses pertaining to other outcomes.

We preregistered our planned research design, hypotheses, data collection, and analyses on the Open Science Framework (Study 1 at https://osf.io/9sr2b/ and Study 2 at https://osf.io/94pks/).

Both studies were approved in the form of a combined application to the Massey University Human Ethics Committee: Northern, application NOR 22/42.

## Study 1: Depression and Anxiety Items

### Study 1 Hypotheses

In Study 1, we investigated the iatrogenic impacts of a survey containing measures of depression and anxiety. We combined both constructs in the same survey because of the strong empirical relationship (and even symptom overlap) between depression and anxiety (see Dobson, 1985; Fried et al., 2016). We compared this survey to a neutral control condition.

Since the existing research suggests that iatrogenic effects are rarely statistically significant and, when present, are small, we hypothesised that iatrogenic effects would either be absent or, if present, their effect sizes would be negligible.

The hypotheses for Study 1 were[2]:

H1.1 Completing sensitive surveys [i.e., with anxiety and depression items] will not produce a higher level of self-reported **distress** than completing neutral control surveys. This was the most central of our hypotheses.

H1.2 Completing sensitive surveys will not produce a higher level of self-reported **tiredness** than completing neutral control surveys. Tiredness was selected as it is a common symptom of both anxiety and depression but has not overtly been investigated in response to sensitive survey completion.

H1.3 Completing sensitive surveys will not produce a lower level of self-reported **happiness** than completing neutral control surveys. We specified happiness as an additional outcome because it can be considered a generic indicator of positive affect.

H1.4 Completing sensitive surveys will not produce a higher level of self-reported **nervousness or anxiety**[3] than completing neutral control surveys. We specified this additional outcome given its connection to the content of the experimental stimuli (which included questions about anxiety).

---

[2] In the manuscript, we have included a prefix for the study number; e.g., hypothesis 1.2 is Study 1, hypothesis 2. These prefixes were not included in the preregistrations themselves, but are included here for clarity.

[3] We retain the original preregistered wording here, but this outcome is perhaps best termed "state anxiety".

## Study 1 Method

### *Participants*

Study 1 used a post-test only, within-subjects design. A within-subjects design was used to maximise statistical power. Given that our preregistered analyses relied heavily on equivalence testing, an *a priori* power analysis was conducted using the TOST (two one-sided tests) equivalence testing spreadsheet (https://osf.io/qzjaj/) by Lakens et al. (2018). In this study, we decided that effect sizes (Cohen's *d*) within the range of Cohen's $d \pm 0.2$ would be considered too small to be practically significant. As the existing literature on iatrogenic effects of assessments for depression and anxiety has little information regarding effect sizes, the selected equivalence bounds are based on the suggested threshold for a small effect size by Cohen (1988). Given these equivalence bounds, the *a priori* power analysis indicated a minimum sample size of 325 participants to achieve 95% power to detect equivalence of effects, assuming a true effect size of zero and an alpha level of 0.05. A target sample size of 342 was thus specified for the study, which included a buffer of 5% (17) to account for potential missing data.

Participants were recruited using convenience sampling via Prolific and were directed to a Qualtrics survey. Using the demographic prescreening function on Prolific, participants were required to have an approval rate of 95% or higher from prior studies to be eligible for participation. Additionally, participants were required to be either Australian or New Zealand residents aged 18 years or older. We selected both countries as they have similar rates of depression and anxiety (see Andrews et al., 2018; Australian Bureau of Statistics, 2022; Wells et al., 2006). Both countries are also predominantly English-speaking and share similar cultural and historical backgrounds, which helped us ensure homogeneity among participants in the experimental and control conditions, and our research team was located across both of these countries at the time of data collection.

Participants received £1.50 for Study 1 (median answering time of 8 minutes and 29 seconds). After applying exclusion criteria (see Table 1), 337 participants remained and were included in the analyses. The application of exclusion criteria is summarised in Table 1, and participant demographic characteristics are displayed in Table 2. For brevity, we include exclusion criteria and demographic characteristics for both studies in the same tables.

**Table 1**

*Participants Excluded From Study 1 and Study 2*

| Exclusion Criteria | *n* | |
|---|---|---|
| | Study 1 | Study 2 |
| Not providing consent | 0 | 2 |
| Below 18 years old | 0 | 0 |
| Living outside of Australia or New Zealand | 4 | 0 |
| Registering any Qualtrics status other than 0 (normal response). This step removed potential survey preview, test, spam and other non-valid responses. | 0 | 0 |
| Not reaching the end of the survey (Finished = 0 in the column as recorded by Qualtrics). | 6 | 6 |
| Failing the first attention check item (see our supplementary materials of Study 1 and Study 2 for attention checks). | 9 | 33 |
| Failing the second attention check item. | 7 | N/A |
| Registering any duplicate responses. Duplicate responses were identified via ProlificID, and only the most recent response from each participant was kept. *This rule only applied after the other exclusion criteria.* | 0 | 1 |
| Total exclusions | 13 | 42 |

*Note*. All exclusion criteria were preregistered.

**Table 2**

*Demographic Characteristics of the Final Samples in Both Studies*

| Characteristics | *n* | |
|---|---|---|
| | Study 1 | Study 2 |
| <u>Gender</u> | | |
| Female | 177 (52.5%) | 524 (60.2%) |
| Male | 155 (46.0%) | 324 (37.2%) |
| Non-binary/gender diverse | 5 (1.5%) | 23 (2.6%) |
| <u>Country</u> | | |
| Australia | 249 (73.9%) | 698 (80.1%) |
| New Zealand | 88 (26.1%) | 173 (19.9%) |
| <u>Age</u> | | |
| 18-24 | 64 (19.0%) | 259 (29.7%) |
| 25-34 | 110 (32.6%) | 315 (36.2%) |
| 35-44 | 87 (25.8%) | 153 (17.6%) |
| 45-54 | 35 (10.4%) | 95 (10.9%) |
| 55-64 | 28 (8.3%) | 34 (3.9%) |
| 65-above | 13 (3.9%) | 15 (1.7%) |
| Overall | *M* = 36.48, *SD* = 13.22 | *M* = 32.56, *SD* = 11.51 |

### *Procedure*

In Qualtrics, participants were randomly assigned to one of two counterbalanced orders: control-then-treatment or treatment-then-control. Study 1 opened and closed on 16 August 2022.

**Treatment condition**. In the treatment condition, the "stimuli" were three measures of depression and anxiety; the Centre for Epidemiological Studies Depression Scale (CES-D: Radloff, 1977), the Kessler Psychological Distress Scale (K10: Kessler et al., 2002), and the Generalised Anxiety Disorder Scale (GAD-7: Spitzer et al., 2006), all within one treatment condition. These measures were selected as they are heavily used in research (including in Australia and New Zealand), are relatively brief, and are available for use in research without payments to their respective publishers.

**Control condition**. The neutral/control survey contained four scales from Hofstee et al.'s (1992) 45 Abridged Big Five Dimensional Circumplex (AB5C) facets which are available from the international personality item pool (Goldberg et al., 2006). These were selected as relatively neutral items that appear unlikely to substantially affect participants' emotions. Please refer to the supplementary materials for Study 1 for a full copy of the survey.

The survey for both conditions comprised 37 items, avoiding confounding by survey length. The stimuli in the two conditions (as described above) were followed by four items measuring the dependent variables. 166 participants (49.3%) completed the treatment condition first, while 171 participants (50.7%) completed the control condition first.

### *Outcome Measures*

Each outcome measure was designed as a *state* measure to ensure that we captured how each participant felt immediately after completing each condition. We used single-item measures to keep the outcome variable section as short as possible, and hence the different conditions as distinct as possible.

The first question asked participants to rate how stressed they felt at the moment on a scale of 0 (not distressed at all) to 10 (the most distressed I have ever felt). This was based on the Distress Thermometer, a single-item measure used clinically to measure distress (Cutillo, 2017).

The second question asked participants to rate their happiness level on a scale of 0 ("the most unhappy I have ever felt") to 10 ("the happiest I have ever felt"). This measure was adapted from Abdel-Khalek's (2006) single-item measure of happiness.

The third question asked participants to rate their tiredness level, on a scale of 0 (not tired at all) to 10 (the most tired I have ever felt), which was adapted from the single-item measure of Severity of Worst Tiredness by Bacci et al. (2017)

The fourth question asked participants to rate their anxiety or nervousness on a scale of 0 (not anxious or nervous at all) to 10 (the most anxious or nervous I have ever felt), which was adapted from the single-item measure of anxiety from the Screening Tool for Psychological Distress (Young et al., 2007).

All responses were indicated on a visual slider ranging from zero to 10. Any slider that was not activated provided a 'no response' in Qualtrics and was not included in the analyses. Each participant completed the outcome measures twice (once after the control condition items, and once after the treatment conditions items).

### *Data Analysis*

We used *R* (R Core Team, 2022) and *RStudio* (Posit Team, 2023) for data analyses. The R packages used were effsize (Torchiano, 2020), TOSTER (Caldwell, 2022; Lakens, 2017), and the tidyverse (Wickham et al., 2019). Paired *t*-tests were conducted to compare the four dependent variable means between the control and treatment conditions. Listwise deletion was used per analysis to remove any cases that did not have responses to the dependent variable across both conditions. Only two data points were missing; these were both for distress levels in the treatment condition.

H1.1 to H1.4 all pertained to the *absence* of effects. Consequently, equivalence tests were used to determine whether the effects were negligible. An equivalence test can reject a null hypothesis that an effect falls within specified equivalence bounds (Lakens et al., 2018). It proceeds by conducting two *one-tailed* tests: One where the null hypothesis is that the true effect is that the true effect falls between -∞ and the lower equivalence bound, and another where the null hypothesis is that the true effect falls between the upper equivalence bound and +∞. If *both* of these null hypotheses can be rejected at the selected alpha level, then there is evidence in favour of equivalence (i.e., that the true effect falls within the equivalence bounds). This procedure has a Type I error rate equal to alpha if the true effect falls exactly on one of the equivalence bounds, and an error rate less than alpha if the true effect is more extreme than that. The procedure described here is also equivalent to examining whether the 90% confidence interval for the parameter of interest falls within the equivalence bounds.

In order for each hypothesis of a null effect to be supported our preregistered inferential criteria required both that there was no statistically significant difference (in a one-tailed[4] test of the null hypothesis of a *zero* effect) and also that the 90% CI for Cohen's *d* was within the equivalence bounds of $-.2 < d < .2$. This comprised a relatively severe test of the hypotheses of absence of effects. The denominator for Cohen's *d* in our calculations was the standard deviation of the paired differences.

The de-identified data, analysis code and questionnaire for Study 1 are available on the Open Science Framework at
https://osf.io/v73t8/?view_only=3ce793a8c0f846758ee33b0683e7b0d2

## Study 1 Results

On average, the participants showed mild to moderate levels of depressive symptoms (*M* = 17.33) on the CES-D (possible range 0 to 60), a moderate level of psychological distress (*M* = 20.82) on the K10 (possible range 10 to 50), and a low level (*M* = 5.54) of anxiety on the GAD-7 (possible range 0 to 21). Post-test means by condition are displayed in Table 3.

**Table 3**

*Posttest Means by Experimental Condition in Study 1*

| Variable | Item | *n* | *Treatment M (SD)* | *Control M (SD)* |
|---|---|---|---|---|
| **Distress** | How distressed do you feel right now? (0 = not distressed at all to 10 = the most distressed I have ever felt) | 335 | 2.16 (2.23) | 2.02 (2.13) |
| **Happiness** | How happy do you feel right now? (0 = the most unhappy I have ever felt to 10 = the happiest I have ever felt) | 337 | 5.36 (1.93) | 5.38 (1.98) |
| **Tiredness** | How tired do you feel right now? (0 = not tired at all to 10 = the most tired I have ever felt) | 337 | 4.56 (2.63) | 4.40 (2.61) |

[4] In hindsight, we probably should not have preregistered one-tailed tests for the null hypotheses of zero effects, since the substantive hypotheses we tested were not directional. We nevertheless retained this criterion in our analyses to avoid deviating from the preregistration. It has the effect of making the tests of our hypotheses (about the absence of effects) especially severe.

| | | | | |
|---|---|---|---|---|
| **State Anxiety** | How anxious or nervous do you feel right now? (0 = not anxious or nervous at all, 10 = the most anxious or nervous I have ever felt) | 337 | 2.61 (2.52) | 2.47 (2.39) |

***Hypothesis 1.1***

A one-sided paired samples *t*-test indicated distress scores were (just) significantly higher in the treatment condition ($M$ = 2.16, $SD$ = 2.23) compared to the control condition ($M$ = 2.02, $SD$ = 2.13); $M_{diff}$ = 0.13, 90% CI [0.01, 0.26], $t$(334) = 1.79, $p$ (1-tailed) = .038, $d$ = 0.10. As such, H1.1 was not supported. However, the 90% CI for the estimated effect size [0.01, 0.19] was within the equivalence bounds of $d$ ± 0.2.

***Hypothesis 1.2***

There was likewise a statistically significant increase in the tiredness scores in the treatment condition ($M$ = 4.56, $SD$ = 2.63) compared to the control condition ($M$ = 4.40, $SD$ = 2.61); $M_{diff}$ = 0.16, 90% CI [0.05, 0.27], $t$(336) = 2.31, $p$ (1-tailed) = .011, $d$ = 0.13, . As such, H1.2 was not supported. The 90% CI for the estimated effect size [0.04, 0.22] was also *not* within the equivalence bounds of $d$ ± 0.2.

***Hypothesis 1.3***

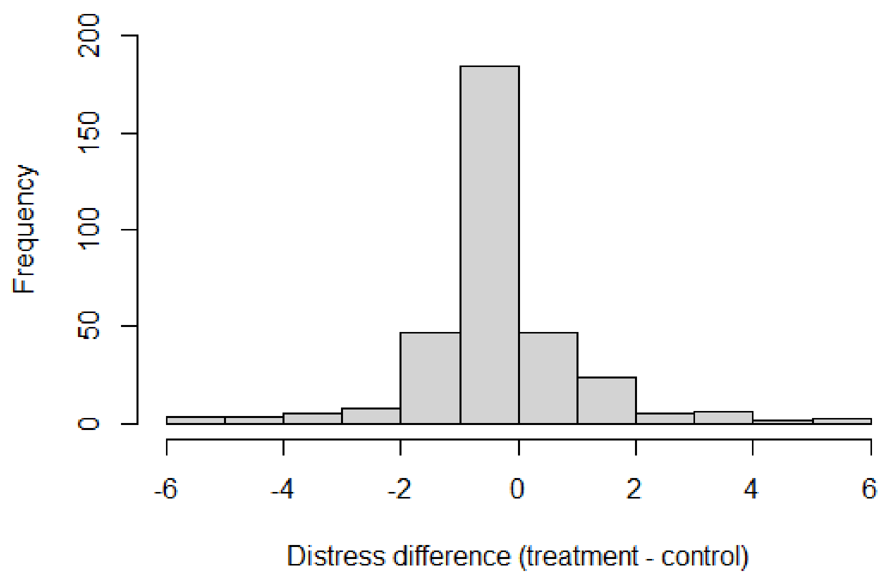
There was no statistically significant difference in happiness scores in the treatment condition ($M$ = 5.36, $SD$ = 1.93) compared to the control condition ($M$ = 5.38, $SD$ = 1.98); $M_{diff}$ = -0.02, 90% CI [-0.17, 0.13], $t$(336) = -0.23, $p$ (1-tailed) = .407, $d$ = -0.01. The 90% CI for the estimated effect size [-0.10, 0.08] was within the equivalence bounds of $d$ ± 0.2. As such, H1.3 was supported.

***Hypothesis 1.4***

There was a statistically significant increase in the state anxiety scores in the treatment condition ($M$ = 2.61, $SD$ = 2.52) compared to the control condition ($M$ = 2.47, $SD$ = 2.39); $M_{diff}$ = 0.13, 90% CI [0.02, 0.24], $t$(336) = 2.02, $p$ (1-tailed) = .022, $d$ = 0.11. As such, H1.4 was not supported. However, the 90% CI for the estimated effect size [0.02, 0.19979] was just within the equivalence bounds of $d$ ± 0.2. A histogram of difference scores is displayed in Figure 1.

**Figure 1**

*Histogram of Differences Scores for Distress, Study 1*




## Study 1 Discussion

Overall, we found statistically significant effects of completing the depression and anxiety items on state distress, tiredness, and anxiety—albeit that for distress and state anxiety there was evidence that the true effects fell within the preregistered equivalence bounds. There was no significant effect on happiness. The picture thus emerging from Study 1 was of the existence of some iatrogenic impacts, albeit very small ones. However, one potential explanation of the statistically significant differences is that the repetition of the outcome measures across two within-subject conditions (treatment, control) might have revealed the hypothesis to some participants, and thus produced demand effects. In addition, Study 1 did not investigate the effect of questions about trauma symptoms.

# Study 2: Trauma Symptoms

## Study 2 Hypotheses

In Study 2 we sought to extend our study of iatrogenic impacts to survey questions about symptoms of trauma. The hypotheses for the second study were:

H2.1 Participants in the trauma survey group will not experience higher **distress** than the non-sensitive survey group.

H2.2 Participants in the trauma survey group will experience their participation to be more **meaningful** than the non-sensitive survey group. To our knowledge, no prior studies have specifically investigated the impacts of completing trauma symptom surveys on perceptions of the meaningfulness of research. We considered it plausible that participants might perceive participating in research on this weighty topic as feeling more meaningful, representing a potential positive emotional outcome.

H2.3 Participants in the trauma survey group will not experience more **reminders of distressing memories** than the non-sensitive survey group[5]. We specified this hypothesis to address the possibility that completing a trauma survey might cause participants to psychologically re-experience trauma.

## Study 2 Method

### *Participants*

In Study 2, recognising the possibility of demand effects, we selected a between-subjects post-test only design. We again preregistered equivalence bounds of $d \pm 0.2$. A power analysis for an equivalence test with an alpha of .05 suggested that 858 participants would be required to achieve 80% power. Due to resource constraints, we elected not to pursue 95% power for this study.

Participants were again recruited via Prolific, with the same inclusion criteria as in Study 1 (residing in New Zealand or Australia and a 95% approval rate on prior studies). We used a

[5] This hypothesis was preregistered, but in the "exploratory analysis" section of the preregistration rather than the "hypotheses" section, reflecting its status as a supplementary hypothesis.

custom blocklist to prevent participants who had taken part in Study 1 from taking part in Study 2, so as to ensure our two studies provide genuinely independent evidence. Participants were paid £0.75 each (median completion time 3m44s). We set a target sample size of 904 in Prolific, including a 5% leeway for exclusions.

The study ran for five weeks to reach the target sample size. After applying the exclusion criteria (see Table 1), our final sample size was $N$ = 871 (see Table 1), above the minimum 858 participants required by the power analysis. Participant demographic characteristics are displayed in Table 2.

## *Procedure*

The first section of the survey enquired about demographic information, which included participants' Prolific ID, age, country of residence and gender. Following that, participants were randomised into two conditions: the trauma survey group (treatment) or the neutral survey group (control) using the randomiser tool in Qualtrics.

Participants in the trauma survey group ($n$ = 441) answered questions from the Trauma Symptom Checklist-40 (TSC-40: Elliott & Briere, 1992). This questionnaire contains 40 items, and participants were asked to rate the frequency of trauma-related symptoms in the past two months on a 4-point Likert-type scale from 0 (never) to 3 (very often). An example symptom is "flashbacks" (sudden, vivid, distracting memories). The TSC-40 assesses trauma-related issues in several domains, including dissociation, anxiety, depression, sexual abuse, sleep disturbance, and sexual problems. It is widely used, as reflected by over 700 citations to Briere and Runtz on Google Scholar as of 12 September 2023. Importantly, the TSC-40 asks only about trauma *symptoms* and not actual experiences of trauma.

Participants in the neutral/control survey group ($n$ = 430) received questions on extraversion (20 items) and openness to experience (20 items). These items were selected from the international personality item pool (Goldberg et al., 2006) and were answered on a 4-point rating scale from 1 (very inaccurate) to 4 (very accurate). We considered these items to be relatively non-sensitive and unlikely to cause distress. Study 2 ran from 17 August 2022 to 24 September 2022, for a duration of approximately 5 weeks and 2 days.

## *Outcome Measures*

In the next section, all participants were asked to answer three questions measuring reactions to the survey. The first question asked participants to indicate how stressed they felt at the moment, indicated on a visual slider scale of 0 ("not at all stressed") to 100 ("the most stress

I have ever felt"). A similar scale was used by Skar et al. (2019) to measure the level of upset after trauma screening.

The second question asked participants the extent to which they agreed that "I found participating in this study personally meaningful". Responses were provided on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). This question originated from the Reactions to Research Participation Questionnaire-Revised (RRPQ-R: Newman et al., 2001), which has been used widely in research on participants' reactions to research.

Finally, the third outcome measure item asked participants to indicate the extent to which they agreed that "The research made me think about things I didn't want to think about". Responses were again provided on a 5-point Likert-type response scale.

### *Data Analysis*

We employed Welch's independent samples t-test to compare the means of outcome variables between the trauma survey and neutral/control survey groups. We opted for Welch's t-test as it obviates the assumption of equal variances that underlie the commonly used Student's t-test (see Delacre et al., 2017).

For H2.1 and H2.3 we again used the two one-sided tests (TOST) procedure to test for equivalence. We again specified equivalence bounds of $d \pm 0.2$. Our preregistered inferential criteria stated that each of these hypotheses would be supported if a two-tailed test of a zero effect was not significant ($p < .05$) *and* the 90% CI for Cohen's $d$ fell within the interval $-0.2 < d < 0.2$.

Study 2 deviated from the preregistration in two ways. First, we intended to conduct an exploratory analysis to see if there was an interaction effect such that the effect of the treatment on distress depends on the severity of trauma symptoms (participants' sum scores on the TSC-40). However, after data collection, we realised the model was impossible to estimate because participants in the control condition did not complete the TSC-40. Second, we planned to retain only the most recent response from each participant with duplicate responses. However, upon further evaluation, we decided to retain the first response in duplicate cases instead because the participants were naive to the study at this stage. Nevertheless, this exclusion criterion only affected one participant with duplicate responses.

The de-identified data, analysis code and questionnaire for Study 2 are available on the Open Science Framework at https://osf.io/qu5xm/?view_only=5600a8aa6f70481ebf771d67558725a9

### Study 2 Results

Participants in the trauma survey condition showed a fairly low mean score ($M$ = 27.97) on the TSC-40 scale (score ranges from 0 to 120). On average, participants reported a low distress level ($M$ = 37.73, $SD$ = 25.47) following their survey participation, were indifferent about the meaningfulness of their participation ($M$ = 3.39, $SD$ = 0.94), were relatively neutral about whether the research reminded them of things they did not want to think about ($M$ = 2.35, $SD$ = 1.21). Posttest means by condition are displayed in Table 4.

**Table 4**

*Posttest Means by Experimental Condition in Study 2*

| Outcome | Item | Trauma *M (SD)* | Control *M (SD)* |
|---|---|---|---|
| Distress | Please indicate how stressed you feel right now, on a scale from 0 (not at all stressed) to 100 (the most stress I have ever felt). | 38.14 (24.52) | 37.30 (26.43) |
| Meaningfulness | I found participating in this study personally meaningful (1 = strongly disagree, 5 = strongly agree) | 3.35 (0.95) | 3.44 (0.93) |
| Unwanted reminders | The research made me think about things I didn't want to think about (1 = strongly disagree, 5 = strongly agree) | 2.75 (1.19) | 1.94 (1.10) |

#### *Hypothesis 2.1*

Welch's independent samples t-test indicated no significant difference in distress between the trauma survey and the neutral survey groups, $M_{diff}$ = 0.84, 90% CI [-2.00, 3.69], $t$(860.37) = 0.486, $p$ (2-tailed) = .627, $d$ = 0.03. The 90% CI for $d$ of [-0.08, 0.14] fell within the equivalence bounds of $d$ ± 0.2. These results support hypothesis 2.1.

### *Hypothesis 2.2*

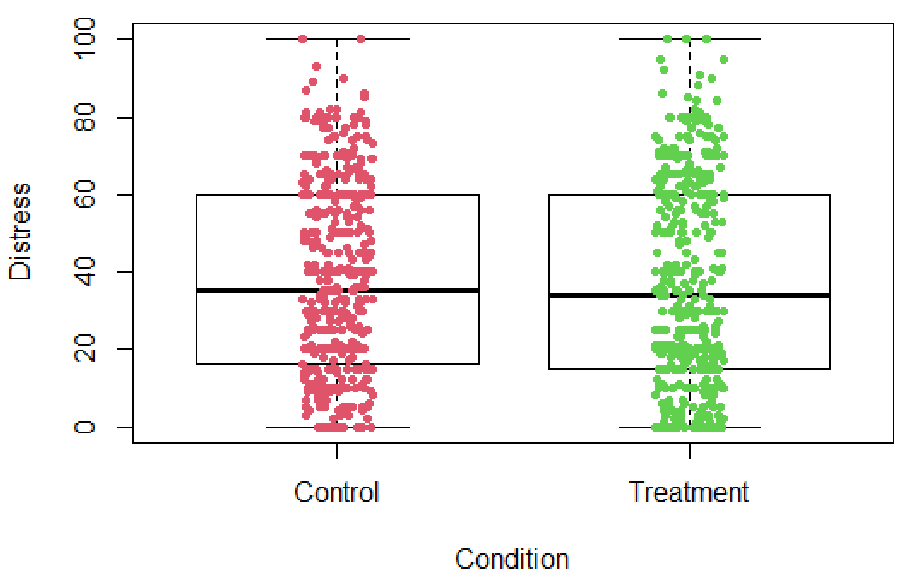
We predicted that participants in the trauma survey group would perceive their survey participation as more meaningful than those in the neutral survey group. However, there was no significant difference in the "meaning" scores between the trauma survey and the neutral survey groups, $M_{diff}$ = -0.086, 90% CI [-0.19, 0.019], $t(868.99) = -1.35$, $p$ (2-tailed) = .177, $d$ = -0.09. Accordingly, hypothesis 2.2 was not supported. Although we did not preregister an equivalence test for this hypothesis (since it involved a directional prediction), an exploratory equivalence test indicated that the 90% CI for $d$ [-0.203, 0.02] was not within the equivalence bounds of $d \pm 0.2$.

### *Hypothesis 2.3*

Participants in the trauma survey group endorsed significantly higher levels of unwanted reminders than in the neutral survey group, $M_{diff}$ = 0.804, $t(866.29) = 10.39$, $p$ (2-tailed) < .001, 90% CI [0.68, 0.93], $d$ = 0.70. The 90% CI for $d$ of [0.59, 0.82] was not within the equivalence bounds of $d \pm 0.2$. These results did not provide support for hypothesis 2.3. Boxplots of distress scores by condition are displayed in Figure 2.

**Figure 2**

*Boxplots of Distress Scores by Condition in Study 2*



## General Discussion

### Summary of Findings

The primary outcome of our studies was state distress. In Study 1, the estimated effect of answering depression and anxiety questions on distress was statistically significant. This meant that our hypothesis (H1.1) of a null effect was not supported according to our preregistered criteria. However, the estimated effect size was very small: a difference of 0.13 on a scale of 0 to 10, or standardised $d$ = 0.13. Its 90% confidence interval also fell within our preregistered equivalence bounds. In Study 2, the estimated effect of answering questions about trauma symptoms on distress was not statistically significant and its 90% confidence interval was within the equivalence bounds, meaning that H2.1 was supported. Overall these findings suggest that sensitive surveys may provoke some non-zero effect on distress, but this effect appears to be *very* small.

We also preregistered hypotheses relating to some additional outcome variables. In Study 1, we found statistically significant effects of answering anxiety and depression questions on state anxiety and tiredness, although in both cases the point estimates of the effect sizes were very small, and for anxiety the estimated effect was significantly within the equivalence bounds. There was no statistically significant effect on happiness, and the estimated effect was also within the equivalence bounds.

In Study 2, we found no evidence of an effect of answering trauma symptoms on the perceived meaningfulness of the research, and the 90% confidence interval for the effect was within the equivalence bounds. However, answers to the item “This research made me think about things I would prefer not to think about” (i.e., unwanted reminders) demonstrated the only genuinely substantial effect in the study: $d$ = .70. However, this expressed dispreference for the thoughts provoked by the study did not appear to translate into an effect on state distress.

Considered in combination, these findings suggest that answering surveys about mental health topics such as anxiety, depression, and trauma symptoms may provoke some negative emotions (e.g., distress, state anxiety). However, these effects appear to be very small, even when measured immediately after the experimental manipulation. Our findings are thus consistent with the growing body of literature suggesting “sensitive” surveys (e.g., on mental health topics) are generally unlikely to have substantial iatrogenic impacts (DeCou & Schumann, 2018; McMurtrie, 2023; Yeater et al., 2012).

## Limitations and Directions for Future Research

First, as convenience sampling was used, our sample was not necessarily representative of the wider Australian and New Zealand populations. That said, our use of Prolific means that participants were drawn from a sampling frame that is particularly likely to engage in research, and is thus of particular salience for meta-psychological research.

Second, we cannot completely rule out the effect of demand characteristics on our results, especially for Study 1 where each participant was exposed to both conditions. It is possible that some participants may have guessed one or more of the hypotheses, and that this could have affected their responses.

Third, the equivalence bounds we specified for equivalence testing were arbitrarily based on Cohen's (1988) definition of a "small" effect size ($d$ = 0.2). Future research could employ empirical methods for determining a smallest effect size of interest, such as the anchor-based method described by Anvari and Lakens (2021).

Fourth, our study only measured outcomes immediately after answering survey items. In future research, it would be useful to include outcome measures at multiple time points to see how quickly iatrogenic impacts (if any) decay over time. On the theme of time, our surveys were relatively short, and it is possible (albeit unlikely) that much longer surveys might provoke more substantial effects.

Fifth, while our Study 2 pertained to trauma, we only studied the effect of answering questions about trauma *symptoms*, and only in a general convenience sample. It is possible that questions that directly ask about trauma, in a sample of participants known to have experienced trauma, might elicit quite different effects.

Finally, we did not record participant ethnicities. It is possible that there are more negative impacts for specific ethnic groups, including Māori (New Zealand's indigenous population), and our samples are unlikely to have included many Māori participants. The Treaty of Waitangi principle of *active protection* (see Te Puni Kōkiri, 2002) means that potential harms of research for Māori are of special salience in New Zealand. It would thus be valuable to design a follow-up study specifically focusing on Māori participants.

## Practical Implications

Given the limitations stated above, and the fact that we did find some evidence of negative iatrogenic impacts (however small), it would be premature for us to conclude that researchers and IRBs can assume that surveys on topics like depression, anxiety, and trauma symptoms will have no distressing effects whatsoever. However, the effects we did find were very small, and this is consistent with the existing research. As such, iatrogenic impacts of completing survey items are unlikely to be a pressing ethical concern in many studies, especially when participants can be adequately informed in advance of the nature of the questions they will answer (facilitating their autonomy to decide whether to take the risk of a distressing effect). Indeed, our research - as is the case for much contemporary research - was conducted with a population who *frequently* complete online surveys, and thus may be well-prepared to judge how they will react to items from various domains.

This said, the risk of harm posed by answering surveys about mental health - even if it is small - is nevertheless a risk of harm, and thus a burden on participants. For a study to meet the ethical principle of justice in the distribution of benefits and burdens of research (Massey University, 2017), participants should receive a benefit from the research that compensates them for any risk of harm, along with the more obvious burden of the time they contribute. Participants on the crowdsourcing platform Mechanical Turk often express frustrations with issues such as spending time on a survey only to be "screened" out, unfair rejections, and poorly-paying studies (Fowler et al., 2022). A cursory examination of the Prolific subreddit (https://www.reddit.com/r/ProlificAc/) demonstrates similar complaints. Participant labour is *work*, and participants deserve to be compensated in a manner that is commensurate with its burdens.

## Conflict of Interest and Funding

This research was supported by funding from the Massey University School of Psychology Postgraduate Research Fund and the Massey University REaDI Fund. Dr Matt Williams is a member of the editorial board at Meta-Psychology.

## Author Contributions

**Conceptualization:** Larry A. Paing, Katie M. P. O'Keefe, and Matt N. Williams.
**Data curation:** Larry A. Paing, Katie M. P. O'Keefe, Brandon L. McMurtrie, and Matt N. Williams.
**Formal analysis:** Larry A. Paing, Katie M. P. O'Keefe, and Matt N. Williams.
**Funding acquisition:** Larry A. Paing, Katie M. P. O'Keefe, and Matt N. Williams.
**Investigation:** Larry A. Paing and Katie M. P. O'Keefe.
**Methodology:** Larry A. Paing, Katie M. P. O'Keefe, and Matt N. Williams.

**Project administration:** Larry A. Paing, Katie M. P. O'Keefe, Brandon L. McMurtrie, and Matt N. Williams.
**Resources:** Larry A. Paing, Katie M. P. O'Keefe, and Matt N. Williams.
**Software:** Larry A. Paing, Katie M. P. O'Keefe, and Matt N. Williams.
**Supervision:** Matt N. Williams.
**Validation:** Larry A. Paing, Katie M. P. O'Keefe, Brandon L. McMurtrie, and Matt N. Williams.
**Visualization:** Matt N. Williams.
**Writing - original draft:** Larry A. Paing, Katie M. P. O'Keefe, Brandon L. McMurtrie, and Matt N. Williams.
**Writing - review & editing:** Larry A. Paing, Katie M. P. O'Keefe, Brandon L. McMurtrie, and Matt N. Williams.

We appreciate the cultural advice provided by Associate Professor Matt Shepherd (Ngāti Tama).

## References


Abdel-Khalek, A. M. (2006). Measuring happiness with a single-item scale. *Social Behavior and Personality: An International Journal*, *34*(2), 139–150. https://doi.org/10.2224/sbp.2006.34.2.139

Akker, O. R. van den, Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S., Valentine, K. D., Veer, A. van 't, & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2625

Andrews, G., Bell, C., Boyce, P., Gale, C., Lampe, L., Marwat, O., Rapee, R., & Wilkins, G. (2018). Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for the treatment of panic disorder, social anxiety disorder and generalised anxiety disorder. *Australian & New Zealand Journal of Psychiatry*, *52*(12), 1109–1172. https://doi.org/10.1177/0004867418799453

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Australian Bureau of Statistics. (2022). *National study of mental health and wellbeing: 2020-2021*. https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/2020-21

Bacci, E. D., DeLozier, A. M., Lin, C.-Y., Gaich, C. L., Zhang, X., Rooney, T., de Bono, S., Hoffman, R., & Wyrwich, K. W. (2017). Psychometric properties of the single-item measure, severity of worst tiredness, in patients with moderately to severely active rheumatoid arthritis. *Health and Quality of Life Outcomes*, *15*(1), 237. https://doi.org/10.1186/s12955-017-0807-5

Barsky, A. J., Saintfort, R., Rogers, M. P., & Borus, J. F. (2002). Nonspecific medication side effects and the nocebo phenomenon. *JAMA*, *287*(5), 622–627. https://doi.org/10.1001/jama.287.5.622

Blades, C. A., Stritzke, W. G. K., Page, A. C., & Brown, J. D. (2018). The benefits and risks of asking research participants about suicide: A meta-analysis of the impact of exposure to suicide-related content. *Clinical Psychology Review*, *64*, 1–12. https://doi.org/10.1016/j.cpr.2018.07.001

Caldwell, A.R. (2022). *Exploring equivalence testing with the updated TOSTER R Package*. PsyArXiv. https://doi.org/10.31234/osf.io/ty8de

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cook, S. L., Swartout, K. M., Goodnight, B. L., Hipp, T. N., & Bellis, A. (2015). Impact of violence research on participants over time: Helpful, harmful, or neither? *Psychology of Violence*, *5*(3), 314–324. https://doi.org/10.1037/a0038442

Courtois, C. A. (2012). Retraumatization and complex traumatic stress: A treatment overview. In M. P. Duckworth & V. M. Follette (Eds.), *Retraumatization: Assessment, Treatment, and Prevention* (pp. 163–190). Routledge.

Cutillo, A. (2017). The distress thermometer: Cutoff points and clinical use. *Number 3 / May 2017*, *44*(3), 329–336. https://doi.org/10.1188/17.ONF.329-336

DeCou, C. R., & Schumann, M. E. (2018). On the iatrogenic risk of assessing suicidality: A meta-analysis. *Suicide and Life-Threatening Behavior*, *48*(5), 531–543. https://doi.org/10.1111/sltb.12368

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Dobson, K. S. (1985). The relationship between anxiety and depression. *Clinical Psychology Review*, *5*(4), 307–324. https://doi.org/10.1016/0272-7358(85)90010-8

Elliott, D. M., & Briere, J. (1992). Sexual abuse trauma among professional women: Validating the Trauma Symptom Checklist-40 (TSC-40). *Child Abuse & Neglect*, *16*(3), 391–398. https://doi.org/10.1016/0145-2134(92)90048-V

Ferrier-Auerbach, A. G., Erbes, C. R., & Polusny, M. A. (2009). Does trauma survey research cause more distress than other types of survey research? *Journal of Traumatic Stress*, *22*(4), 320–323. https://doi.org/10.1002/jts.20416

Fowler, C., Jiao, J., & Pitts, M. (2022). Frustration and ennui among Amazon MTurk workers. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01955-9

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320. https://doi.org/10.1016/j.jad.2015.09.005

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Gould, M. S., Marrocco, F. A., Kleinman, M., Thomas, J. G., Mostkoff, K., Cote, J., & Davies, M. (2005). Evaluating iatrogenic risk of youth suicide screening programs: A randomized controlled trial. *JAMA*, *293*(13), 1635–1643. https://doi.org/10.1001/jama.293.13.1635

Hahn, R. A. (1997). The nocebo phenomenon: Concept, evidence, and implications for public health. *Preventive Medicine*, *26*(5), 607–611. https://doi.org/10.1006/pmed.1996.0124

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146–163. https://doi.org/10.1037/0022-3514.63.1.146

Jaffe, A. E., DiLillo, D., Hoffman, L., Haikalis, M., & Dykstra, R. E. (2015). Does it hurt to ask? A meta-analysis of participant reactions to trauma research. *Clinical Psychology Review*, *40*, 40–56. https://doi.org/10.1016/j.cpr.2015.05.004

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., Degenhardt, L., de Girolamo, G., Dinolova, R. V., Ferry, F., Florescu, S., Gureje, O., Haro, J. M., Huang, Y., Karam, E. G., Kawakami, N., Lee, S., Lepine, J.-P., Levinson, D., … Koenen, K. C. (2017). Trauma and PTSD in the WHO world mental health surveys. *European Journal of Psychotraumatology*, *8*(sup5), 1353383. https://doi.org/10.1080/20008198.2017.1353383

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Lakeman, R., & FitzGerald, M. (2009). The Ethics of Suicide Research. *Crisis*, *30*(1), 13–19. https://doi.org/10.1027/0227-5910.30.1.13

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Massey University. (2017). *Code of ethical conduct for research, teaching and evaluations involving human participants*. https://www.massey.ac.nz/massey/research/research-ethics/human-ethics/code-ethical-conduct.cfm

McMurtrie, B. L. (2023). On the iatrogenic effect of sensitive-topics survey research: A scoping review. *Journal of European Psychology Students*, *14*(1), 39–53. https://doi.org/10.59477/jeps.601

Newman, E. (2008). Assessing trauma and its effects without distress: A guide to working with IRBs. *APS Observer*, *21*. https://www.psychologicalscience.org/observer/assessing-trauma-and-its-effects-without-distress-a-guide-to-working-with-irbs

Newman, E., Willard, T., Sinclair, R., & Kaloupek, D. (2001). Empirically supported ethical research practice: The costs and benefits of research from the participants' view. *Accountability in Research*, *8*(4), 309–329. https://doi.org/10.1080/08989620108573983

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400–424. https://doi.org/10.1111/j.1745-6924.2008.00088.x

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Paulhus, D. L., & Vazire, S. (2009). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 224–239). Guilford Press.

Pennebaker, J. W. (1982). *The psychology of physical symptoms*. Springer-Verlag.

Posit team (2023). *RStudio: Integrated development environment for R* (Version 2023.06.2) [Computer software]. http://www.posit.co/

R Core Team (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computing language and environment]. https://www.R-project.org/

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Rinehart, J. K., Nason, E. E., Yeater, E. A., & Miller, G. F. (2017). Do some students need special protection from research on sex and trauma? New evidence for young adult resilience in "sensitive topics" research. *The Journal of Sex Research*, *54*(3), 273–283.

https://doi.org/10.1080/00224499.2016.1156047

Robinson, J., Pan Yuen, H., Martin, C., Hughes, A., Baksheev, G. N., Dodd, S., Bapat, S., Schwass, W., McGorry, P., & Yung, A. R. (2011). Does screening high school students for psychological distress, deliberate self-harm, or suicidal ideation cause distress—And is it acceptable? An Australian-based study. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *32*(5), 254–263. https://doi.org/10.1027/0227-5910/a000087

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, *16*(4), 376–390. https://doi.org/10.20982/tqmp.16.4.p376

Skar, A.-M. S., Ormhaug, S. M., & Jensen, T. K. (2019). Reported levels of upset in youth after routine trauma screening at mental health clinics. *JAMA Network Open*, *2*(5), e194003–e194003. https://doi.org/10.1001/jamanetworkopen.2019.4003

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Te Puni Kōkiri. (2002). *The principles of the Treaty of Waitangi as expressed by the courts and the Waitangi Tribunal*. https://www.waitangitribunal.govt.nz/treaty-of-waitangi/principles-of-the-treaty/

Torchiano, M. (2020). *effsize: Efficient effect size computation* (version 0.8.1) [Computer software]. https://CRAN.R-project.org/package=effsize

Wells, J. E., Oakley Browne, M. A., Scott, K. M., McGee, M. A., Baxter, J., & Kokaua, J. (2006). Prevalence, interference with life and severity of 12 month DSM-IV disorders in Te Rau Hinengaro: The New Zealand mental health survey. *Australian & New Zealand Journal of Psychiatry*, *40*(10), 845–854. https://doi.org/10.1080/j.1440-1614.2006.01903.x

Wickham. H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open*

*Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, M. N., Bååth, R. A., & Philipp, M. C. (2017). Using Bayes factors to test hypotheses in developmental research. *Research in Human Development*, *14*(4), 321–337. https://doi.org/10.1080/15427609.2017.1370964

World Health Organization. (2022). *Mental disorders*. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Yeater, E., Miller, G., Rinehart, J., & Nason, E. (2012). Trauma and sex surveys meet minimal risk standards: Implications for institutional review boards. *Psychological Science*, *23*(7), 780–787. https://doi.org/10.1177/0956797611435131

Young, Q.-R., Ignaszewski, A., Fofonoff, D., & Kaan, A. (2007). Brief screen to identify 5 of the most common forms of psychosocial distress in cardiac patients: Validation of the screening tool for psychological distress. *Journal of Cardiovascular Nursing*, *22*(6), 525. https://doi.org/10.1097/01.JCN.0000297383.29250.14