



Meaningful change definitions: sample size planning for experimental intervention research

Stefan L. K. Gruijters & Gjalt-Jorn Y. Peters

To cite this article: Stefan L. K. Gruijters & Gjalt-Jorn Y. Peters (2020): Meaningful change definitions: sample size planning for experimental intervention research, *Psychology & Health*, DOI: 10.1080/08870446.2020.1841762

To link to this article: <https://doi.org/10.1080/08870446.2020.1841762>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 19 Nov 2020.



Submit your article to this journal [↗](#)



Article views: 771





View related articles [↗](#)



View Crossmark data [↗](#)

Meaningful change definitions: sample size planning for experimental intervention research

Stefan L. K. Gruijters^a  and Gjalt-Jorn Y. Peters^{b,c} 

^aFaculty of Psychology, General Psychology, Open University of the Netherlands, Heerlen, the Netherlands; ^bFaculty of Psychology, Methodology & Statistics, Open University of the Netherlands, Heerlen, the Netherlands; ^cFaculty of Psychology and Neuroscience, Work and Social Psychology, Maastricht University, Maastricht, the Netherlands

ABSTRACT

Experimental tests of interventions need to have sufficient sample size to constitute a robust test of the intervention's effectiveness with reasonable precision and power. To estimate the required sample size adequately, researchers are required to specify an effect size. But what effect size should be used to plan the required sample size? Various inroads into selecting the *a priori* effect size have been suggested in the literature—including using conventions, prior research, and theoretical or practical importance. In this paper, we first discuss problems with some of the proposed methods of selecting the effect size for study planning. We then lay out a method for intervention researchers that provides a way out of many of these problems. The proposed method requires setting a meaningful change definition, it is specifically suited for applied researchers interested in planning tests of intervention effectiveness. We provide a hands-on walk through of the method and provide easy-to-use *R* functions to implement it.

ARTICLE HISTORY

Received 5 November 2019
Accepted 17 October 2020

KEYWORDS

Effect size; smallest effect size of interest; sample size planning; intervention research; meaningful change definitions; practical significance

Awareness of the importance of having an adequate sample size for power and precision has spread through empirical psychology, but actually estimating the required sample size in a meaningful way is often perceived as challenging. The reason for this mostly relates to one aspect: one needs to decide on what effect size magnitude to use for planning the required sample size, because the required sample size is dependent on the effect size of an intervention. Specifying this effect size *a priori* can be a challenging activity for researchers, arguably 'the most difficult part of power analysis' (Cohen, 1992, p. 156). In this paper, we introduce a method for determining

CONTACT Stefan L. K. Gruijters  mail@stefangruijters.nl  Faculty of Psychology, General Psychology, Open University of the Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, the Netherlands. All scripts, functions and other materials described in this article are publicly available at the Open Science Framework (<https://osf.io/xrs23/>). A pre-print of this paper has been posted previously to PsyArxiv (<https://psyarxiv.com/jc295/>).

© 2020 Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

what effect size magnitude to use for planning sample size of experimental intervention research. The method we propose involves backwards engineering a definition of meaningful change, as it might result from an intervention, ultimately arriving at a standardized effect size (e.g., Cohen's d) denoting the smallest effect size one is interested in to discover. Lakens (2014) uses the label 'smallest effect size of interest' (SESOI) for the smallest effect size that is of interest to a researcher. The proposed method in this paper specifically enables applied researchers to specify such a SESOI – which, in turn, can be used to plan the required sample size for a given experimental test of an intervention.

This paper is organized as follows. First, we discuss and evaluate previously proposed 'gold standard' strategies to select a SESOI for power or precision analysis. Second, we describe a method to arrive at *meaningful change definitions*, which can be used to determine the SESOI in experimental tests of interventions. We illustrate through several steps how the proposed method to determine the SESOI can be used, accompanied by functions in the *R*-software package 'behaviorchange' (Peters, 2020a)

What effect size should be used to plan sample size?

The question which SESOI to use for sample size planning is an important first-step for any research project. Before discussing tactics to make inroads into this issue, we delve a bit deeper into some common – but not very optimal – ways of dealing with this question. One strategy to choose a SESOI relies on using rules of thumb for effect size thresholds that were suggested in earlier publications, such as Cohen's (e.g., 1962, 1988) benchmarks of "small" ($r > .1$), "medium" ($r > .3$), and "large" ($r > .5$). Research proposals, then, sometimes include a statement in line with: "this project aims to be able to detect at least a medium size effect, so the power analysis and corresponding sample size estimates are based on Cohen's $d = .5$ ". The (often implicit) rationale for this decision is likely that it is believed that a "medium" effect size is something worth to detect, as opposed to effect size magnitudes that are labelled "small". This choice hinges on the circular argument that things labelled as "small" are not interesting findings because they are small, and that effects of medium and larger size *are* interesting because they are not small.

Cohen (1962, 1988) understood that the justification for his cut-offs was tenuous. The benchmarks were not intended to be used as a key method for a power analysis or sample size planning. Instead, Cohen suggested them as a last resort when no other guidance is available. As he put it: '... these proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition' (Cohen, 1988, p. 532). The benchmarks were based on the intuition that some range of ES magnitudes (labelled medium) would become perceptible, 'visible to the naked eye of a careful observer' (Cohen, 1992, p. 156). Hence, the labels "small", "medium" and "large" do not carry any inherent meaning related to the *importance* or *triviality* of an effect magnitude – though, they are often interpreted as if they do. As Cortina and Landis (2009) put it: 'reflexive dismissal of small effect sizes by researchers

reflects an urban (and for that matter, rural) legend that small effect size values always mean the same thing and justify labels such as weak effect or trivial effect' (p. 288).

Illustrations of the falsity of this myth abound (Aguinis & Harden, 2009; Cortina & Landis, 2009). For example, Furukawa and Leucht (2011) point out that improving a depression treatment (e.g., improving an anti-depressant) by a Cohen's d of 0.2 could, given the incidence of depression in Japan, 'bring about remission in additional 100 thousand or more people that would not have done so on the current treatment ($p.5$)'. It is unclear by what substantive standards such results are reasonably labelled as "small". Rosenthal and Rubin (1982) sketched a (hypothetical) scenario in which a correlation of $r = .1$ between treatment and outcome would correspond to a 10% decrease of some discrete outcome incidence (e.g., illness rate), in what they referred to as a binomial effect size display (BESD; Rosenthal, 1990, 1991). These examples make clear that "small" or "large" (or anything in between) effect size magnitudes are relative concepts and fully dependent on the context of the research. Because of this, general rules of thumb are unlikely to be calibrated to the context of any specific study.

Ways forward to specifying a meaningful SESOI

Cohen et al. (2003) specify three general approaches of choosing a SESOI to be used for sample size planning – and note that these are not on equal footing in terms of quality. In abbreviated form, strategies suggested to use for planning sample size are as follows: 1) use the observed effect size of previous studies, 2) base the effect size parameter on theoretical or practical significance, and 3) use conventions. As discussed, conventions are largely unfounded and not calibrated to any specific study's context, and are therefore best avoided. This implies that researchers' preference instead should move towards using previous studies, or theoretical or practical significance to derive an effect size.

The first strategy, to use an observed effect size of one or more previous studies as the SESOI, was aimed at close (or direct) replications. As Cohen et al. (2003) put it: 'to the extent that studies that have been carried out [...] are closely similar to the present investigation, the [effect sizes] found in these studies reflect the magnitude that can be expected' (p. 52). At first blush, this strategy appears sensible for replication attempts of previous work; giving both exact and conceptual replication studies a means to specify the SESOI and plan the required sample size. However, when there is insufficient conceptual overlap (e.g. changes in design or operationalisation), this rationale for the SESOI dissipates entirely – because there is no reason to assume that the conceptual differences are orthogonal to the studied effect. In many cases, it is likely not feasible to determine with reasonable certainty whether sufficient conceptual overlap between an earlier and current study exists (see also Lakens & Evers, 2014). Such lack of clarity on study comparability is a cogent argument for avoiding the strategy. Additionally, even if one aims to conduct a direct replication, there are (at least) three more reasons to refrain from basing sample size planning procedures on effect size estimates from earlier studies.

First, if only one study has been conducted, given the small sample sizes and low statistical power that have long been customary in the psychological literature

(Marszalek et al., 2011), the effect size estimate obtained in that study likely originates from an exceedingly wide sampling distribution. In other words, the observed effect size in that single study is to a large degree arbitrary, may differ considerably in the next direct replication (Peters & Crutzen, 2020), and as such does not provide a solid starting point for inferences about the likely population effect size (see also Anderson & Maxwell, 2017). If multiple studies have been conducted and their effect size estimates are extracted from the extant literature, meta-analysis may yield a more reliable estimate. Unfortunately, publication bias hinders this route as well. Publication bias is the phenomenon that journals are unlikely to accept null or negative findings (Fanelli, 2012), which in turn discourages researchers from submitting such studies in the first place. This means that the effect sizes in published studies exhibit an upward bias of unknown magnitude (e.g., Driessen et al., 2015). Methods to correct for uncertainty in the effect size estimate and publication bias have, however, been described in the literature (e.g., Anderson et al., 2017; Du & Wang, 2016). Nonetheless, even if unbiased estimates of the population effect size can be obtained, not every non-zero effect size is necessarily worthwhile to examine. For example, a costly treatment may have shown an effect size of $d = .3$ in a meta-analysis of earlier studies that did not show evidence of publication bias; but that effect size may simply be uninteresting if it fails to offset the treatment costs by a considerable margin.

As a solution to these problems, some researchers instead resort to estimating the effect size based on pilot studies. However, researchers are urged caution regarding the use of pilot studies to guide power calculations for study proposals (Kraemer et al., 2006). Pilot studies are typically designed to detect problems in a study's proposed procedure. To this end, relatively small samples often suffice, so applying the appropriate sample size computations for pilots (see Viechtbauer et al., 2015) often yields sample sizes that do not allow reasonably accurate estimation of parameters. As a result, the observed effect size in a pilot study will often deviate considerably from the population effect size. To acquire a reasonably precise estimate of the population effect size, a pilot study would need to have sufficient precision – and therefore a large sample size. This, of course, defeats the whole purpose of conducting a pilot study (see also Lakens, 2014; Lakens & Evers, 2014).

Given the discussed limitations of relying on conventions or expectation, the third proposed strategy described by Cohen et al. (2003) – practical or theoretical significance – provides a remaining inroad into selecting the SESOI. Basing the SESOI on practical or theoretical significance is from our vantage point the first that needs to be attempted by researchers, and therefore preferred over Cohen et al. (2003) alternatives. Simply put, this is because importance supersedes expectation or convention – it is not clear why researchers would spend finite resources such as money, time, and incurred participant burden on examining an association based on expectation when the expected magnitude is thought to be practically or theoretically irrelevant. Additionally, it may be worthwhile to invest resources in examining whether a treatment reaches a Cohen's d of 0.2 (see the example in the previous section) based on the expected number of people this treatment would impact – even though this may be a small effect by convention.

However, basing the SESOI on theoretical importance is utopic by our estimation – at least given the present state of theoretical psychology (see also Lakens, 2014; Muthukrishna & Henrich, 2019; Smaldino, 2017). A prerequisite for this strategy is that theories are specified with sufficient mathematical detail to allow a meaningful expression of importance. For example, formal theory could predict that under circumstance z (e.g., an increase of p percent on variable x) variable y increases by q percent on average. This magnitude could then be used to as the effect size parameter in a power or precision analysis. While such precision in formal theory may be abundant in the sciences, social science (thus far) is mostly involved with testing informal theory lacking such precision (see also Meehl, 1967). The majority of theories are specified only in terms of predicting the presence of associations (non-zero associations), rather than making predictions on the magnitude of those associations. For this reason, most social science theories do not provide a solid springboard to determine what effect size magnitudes are theoretically interesting, and which are not.

Specifying the SESOI based on practical importance

A strategy allowing specification of the SESOI, one particularly feasible in the context of applied intervention research, is to consider practical importance. At this point it is helpful to make a pragmatic division between two types of outcome measures. Whereas some studies examine associations on measures which have (at least, by hypothesis) observable consequences in the ‘real world’ (on practically meaningful scales), other projects aim to draw conclusions based on measures that are abstracted from the real world (practically non-meaningful scales). For many basic science projects, there is no direct connection to practically meaningful units allowing the determination of a SESOI. For example, a factor could cause changes in individuals’ attitude toward a behaviour (measured on a 7-point Likert scale) by a given Cohen’s d , but what is the minimum amount of change in Likert scaled attitudes that is worthwhile to study?

Clinical epidemiologists have long recognized the importance of finding out what changes on a non-meaningful measure can be seen as a ‘meaningful’ intervention result (e.g., Guyatt et al., 2002; Jaeschke et al., 1989; McGlothlin & Lewis, 2014). As put by Guyatt et al. (2002): ‘If a patient with chronic lung disease improves by 5 points in physical function, will she now be able to climb a flight of stairs comfortably, keep up with her spouse when they go for a walk, and resume playing with her grandchildren? Or will she remained incapacitated by exertional dyspnea?’ (p. 373). In this literature, the change on some unstandardized or standardized outcome needed to create a meaningful change for patients or clients is known as the minimal (clinically) important difference (e.g., Jaeschke et al., 1989; McGlothlin & Lewis, 2014). Establishing such a minimal important difference is usually done with one of the following three methods (see McGlothlin & Lewis, 2014). First, using the *consensus method*, which relies on a panel of experts to determine what amount of quantitative change is clinically meaningful. Second, using a *distributional method*, which involves looking at statistical properties (score distributions) of the quantitative outcome to determine what constitutes meaningful change. Third, by using an *anchor-based method*, which relates

changes on a quantitative outcome to an independent qualitative measure of improvement. For instance, results of clinical and health interventions can often be anchored in noticeable improvement in quality-of-life (QOL) markers, such as ‘can walk the stairs again’ in the lung disease example above.

Anvari and Lakens (2019) recently described a psychological application of the anchor-based approach in more depth, suggesting methods to base the SESOI on the concept of a minimally detectable difference – ‘the smallest effect size that is associated with a subjectively noticeable change at the individual level’ (p. 1). To illustrate their method, consider a study examining the benefit of a cognitive treatment in reducing rumination symptoms of anxiety. In this example, one external QOL criterion could be sleep quality, as a qualitative measure of improvement. That is, one could examine how much rumination scores need to change in order for individuals to experience a transition from ‘not feeling rested’ towards ‘feeling rested’. Using the anchoring method enables researchers to estimate the threshold Cohen’s d that corresponds to a meaningful change on such a QOL indicator. It could, for example, turn out that an intervention of Cohen’s d of 0.2 does not result (on average) in any changes in the subjective judgment of feeling rested, whereas a Cohen’s d of 0.25 does. This would be an indication to specify the SESOI at $d=0.25$, and plan sample size accordingly.

In addition to QOL-markers, there are various conceivable external criteria that could serve as qualitative anchors. Overall, we think the anchoring method, as applied by for instance Anvari and Lakens (2019) on a measure of affect, is one example of the way forward to specifying a meaningful SESOI in intervention studies where the outcome measures are not practically meaningful – putting researchers in the position to plan their sample sizes for power or precision accordingly. We aim to describe a different method, suitable for applied research projects that allow importance to be defined on the primary intervention outcome directly. The strategy we propose to estimate the SESOI is based on setting a *meaningful change definition* (MCD) to define meaningful change on a continuous outcome variable. We use a concrete example in what follows to provide a step-by-step walk-through of the MCD-method.

Meaningful change definitions: a case-study using a physical activity intervention

Consider an intervention that is designed to increase physical activity in a population, by trying to influence several psychological determinants of the behaviour (e.g., people’s motivation to exercise, attitudes, and self-efficacy). The recommended development process for such interventions generally includes a number of pre-tests, to optimize the effect of specific intervention components on the targeted determinants, as well as an evaluation of the intervention as a whole (Bartholomew Eldregde et al., 2016). Such studies often do not measure intervention effects directly on a dichotomous measure of success (e.g., physical activity status after intervention; ‘sufficient’ versus ‘not sufficient’). Rather, researchers interested in experimentally testing interventions often rely on continuous (scaled) outcome measures – and there are, of course, good statistical arguments to do so (Cohen, 1983; DeCoster et al., 2009;

MacCallum et al., 2002). Nonetheless, to establish a meaningful SESOI, interventionists using scaled outcomes need to have some idea of what range of outcome scores count as ‘desirable’ and ‘undesirable’ (or at least, ‘insufficient’). Indeed, adequate intervention development requires that such goals are specified *in advance* of intervention tests with sufficient detail (Bartholomew Eldregde et al., 2016).

There are various conceivable ways to set meaningful intervention goals. One example is a consideration of cost-benefit: if the costs are less than benefits (according to some definition) then the effect size magnitude may be worthwhile to study. For example, in health care research the effectiveness of a given intervention is often quantified in terms quality-adjusted life years (QUALY) that are gained due to intervention. In other settings, the financial costs per participant exposed to the intervention may be known and could be used to set an MCD. However, for many health-related interventions the desired outcome is difficult to quantify in terms of QUALY or concrete financial cost-benefits. A remaining option, then, is to determine the MCD based on a consensus method – the research team determines in accordance with stakeholders and policy-makers feasible and desirable intervention goals. Since definitions of meaningful change are subject to policy, financial, and societal and cultural considerations – and as such, usually beyond the purview of the primary research team – we do not consider it feasible to provide context-independent guidelines on how to set a given MCD.

STEP 1: Specifying a threshold definition

All scripts, functions and other materials described in this section are available at the Open Science Framework (<https://osf.io/xrs23/>). The MCD-method is implemented in the R-package ‘behaviorchange’ (Peters, Crutzen & Gruijters, 2020a). Additionally, the method is also available in the ‘behaviorchange’ JAMOVI module (<https://www.jamovi.org/>).

For an intervention designed to promote physical activity levels, the MCD-procedure requires that some external criterion is used to specify the physical activity levels that are considered ‘positive’. In this instance, we use ‘expert consensus’ as an external criterion: both the U.S. Department of Health and Human Services and the Dutch health council (Gezondheidsraad, 2017; U.S. Department of Health & Human Services, 2018) maintain a policy recommending a minimum of 150 minutes of weekly exercise (moderate-intensity). In this example, this consensus criterion will serve as the project’s *threshold definition* (TD). Cases (i.e., participants) that meet such a criterion are referred to as ‘positive events’, and those not meeting such levels are labelled as ‘negative events’. Using this TD on the continuous ‘minutes of weekly exercise’ outcome, one can distinguish two types of events: ‘positive events’ (> 150 minutes of weekly exercise) and ‘negative events’ (< 150 minutes of weekly exercise).

STEP 2: Estimating the control event rate corresponding to a threshold definition

Once a TD has been specified (and included in the preregistration plan) one needs to estimate the base-rate occurrence of this outcome – that is, what proportion of the

Table 1. Required sample sizes to estimate a base-rate or control event rate as a function of the desired half-widths of a 95% confidence interval.

CER	95% confidence interval half-width				
	0.01	0.025	0.05	0.1	0.15
0.01	496	112	43	18	11
0.05	1926	333	93	28	14
0.10	3556	592	157	43	20
0.25	7300	1190	305	79	36
0.50	9700	1573	401	103	46

Note. Base-rate values depict proportions.

population currently meets this standard? In the clinical epidemiology literature (e.g., Furukawa & Leucht, 2011), this value is also commonly referred to as the control event rate (CER). Conversely, the event rate in the intervention or experimental group is often referred to as the experimental event rate (EER). In some instances, it is possible that such CER population information is publicly available: for instance, in the Netherlands we know that in 2018 roughly 47% of the general Dutch population over 18 years of age meets the TD (>150 minutes) for physical exercise (CBS/RIVM., 2018). If current population data are not available, the CER would need to be estimated using sample data. In these instances, researchers would need to estimate the CER value before planning an experimental test of the intervention. For instance, take the Dutch situation, but assume that the ambitious researchers in this example have decided that the intervention goal is to get more people to exercise at least 160 minutes a week (instead of 150 minutes). In that case, the base-rate is again unknown, and needs to be estimated. In such cases, it is important to consider the degree of uncertainty accompanying a particular estimate of the population CER. When random samples are used to estimate the CER, it is possible to account for uncertainty (random sampling error) by obtaining a confidence interval for the CER estimate and involving the interval in the MCD-method (see Step 4 on how to do so).

The required sample size to estimate a proportion can be computed with Accuracy In Parameter Estimation (AIPE) procedures (e.g., Cumming, 2014; Kelley & Rausch, 2006; Lai & Kelley, 2012). For example, in *R-software* by using the 'ufs' (Peters, 2020b) or 'MBESS' (Kelley, 2020) package. Table 1 shows the required sample sizes for estimating proportions of .01, .05, .1, .25, and .5, with 95% confidence interval half-widths of .01, .025, .05, .1, and .15. Note that in the context of intervention evaluations, such data will often have to be collected as part of the needs assessment; because an analysis of the scope of the problem is often a prerequisite to allocating resources to intervention development in the first place.

STEP 3: Deciding on a meaningful change definition

Once both the TD (> 150 minutes of exercise) and corresponding CER (47%) is specified, the next step is to specify the smallest success rate difference that is considered meaningful – that is, specifying a meaningful change definition (MCD). Similar to our TD, such an MCD needs to be based on external criteria, such as extant policy, based on expert-consensus or cost-benefit considerations. In the current illustrative case-study, it is assumed that given the costs of the hypothetical intervention a change of

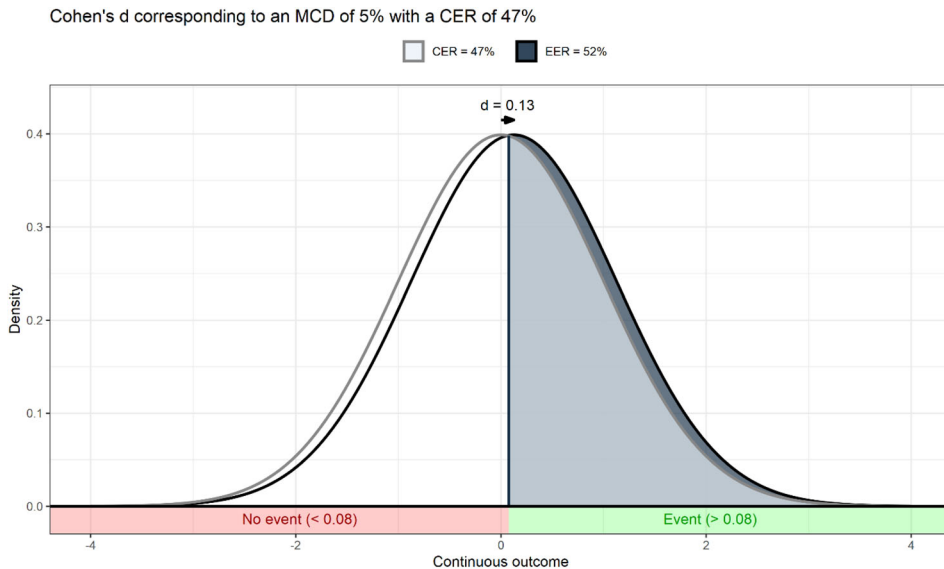


Figure 1. An illustration of the MCD-method to estimate a SESOI. The graph depicts the required Cohen's d value required to increase the percentage of positive events by 5% for a behaviour with a base-rate (CER) of 47%.

5% is considered practically meaningful (i.e. cost effective). Thus, the MCD of 5% would correspond in this situation to a desired behaviour change in this population from a CER= 47% to EER = 52%.

This increase of 5% can also be expressed as an absolute frequency: in the Netherlands in 2019, around 14 million people were 18 years or older (Centraal Bureau voor de Statistiek, 2019); therefore, 47% corresponds to about 6.5 million, and an improvement of 5% would mean that give or take 700.000 more would meet the threshold – in the hypothetical scenario that the intervention would target this entire population. Thus, the TD and MCD together provide a convenient interface between frequently used outcome measures in intervention research and measures that more familiar to politicians, policy makers, practitioners and members of the general public. More importantly for the purposes of this paper, they enable computing the corresponding SESOI.

STEP 4: Estimating the SESOI based on the MCD

Under parametric assumptions it is possible to estimate with a simple equation the standardized difference (Cohen's d value) that corresponds to an MCD of 5% (given CER= 47%). The equation is as follows,

$$d = \Phi^{-1}(CER + MCD) - \Phi^{-1}(CER) \quad (1)$$

where Φ^{-1} denotes the inverse of a standard normal distribution. Equation 1 estimates the Cohen's d value that would be needed to increase a given base rate of desirable outcomes (e.g., CER= 0.47) by a given MCD (e.g., an additional 5% positive events due

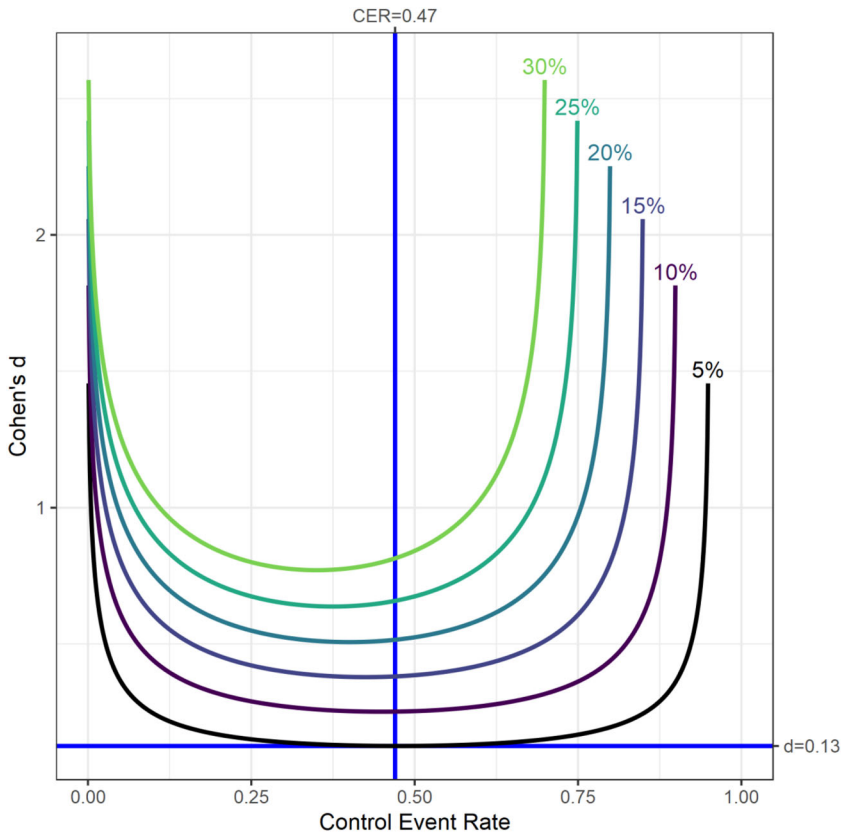


Figure 2. The relationship between the control event rate (CER), Cohen's d and the MCD. The intersection of the dark blue lines depict the estimated SESOI for the exercising example used in this paper. The different lines represent different MCD-values (ranging from 5% to 30%).

to intervention). This function is implemented in the *R*-package 'behaviorchange' as 'dMCD', and only requires the specification of the CER (0.47) and the MCD (0.05).

```
install.packages('behaviorchange');
behaviorchange::dMCD(cer=.47, mcd=.05);
```

Entering these numbers in the dMCD function results in the following: in order to achieve an MCD of 5% given CER=.47, a Cohen's $d=0.13$ would be sufficient to obtain this change (see Figure 1). In case of uncertainty associated with the CER value (see also Step 2) it is possible to include a sensitivity analysis to gauge how variance in the CER affects the result of the MCD-procedure. Lower and upper bounds of the CER 95% confidence interval would be natural candidates. For instance, assuming that the above-used CER = 0.47 is a sample estimate with a confidence interval ranging from [0.44; 0.50], we obtain Cohen's d values for the lower bound, point estimate and upper bound as follows:

```
behaviorchange::dMCD(cer=c(.44, .47, .50), mcd=.05);
```

Note that it is also possible to estimate the required SESOI (d) when the estimate of the CER is derived from distributions other than Gaussian. For example, outcome

Table 2. The Cohen's d corresponding to, and required sample size for, a variety of Control Event Rates and Meaningful Change Definitions, for both Null Hypothesis Significance Testing and Accuracy In Parameter Estimation approaches.

CER	MCD	d	N (80% power)	N (95% power)	N for 95% CI with half-width of $d = .10$	N for 95% CI with half-width of $d = .25$
0.05	0.025	0.21	714	1182	1546	248
0.05	0.050	0.36	246	404	1562	250
0.05	0.100	0.61	88	142	1609	258
0.05	0.250	1.12	28	44	1778	285
0.25	0.025	0.08	4908	8124	1538	247
0.25	0.050	0.15	1398	2314	1541	247
0.25	0.100	0.29	376	620	1553	249
0.25	0.250	0.67	72	118	1623	260
0.50	0.025	0.06	8724	14442	1538	246
0.50	0.050	0.13	1860	3078	1540	247
0.50	0.100	0.25	506	834	1549	248
0.50	0.250	0.67	72	118	1623	260

Note. CER = control event rate, MCD = meaningful change definition, CI = confidence interval, d = Cohen's d estimate for the standardized mean difference.

variables may be left or right skewed in a population. For some distributions exhibiting right skewness, the quantiles can for instance be estimated using a lognormal distribution to derive d . In the 'behaviorchange' package it is possible to change the default distribution parameter (Φ) and estimate the required d using user-specified distributions (such as lognormal, beta, and so forth). This is implemented in the dMCD function under the 'dist' parameter. Before implementing the MCD-method, it is important to consider what distribution best describes the continuous variable in the population. For example, variables with ratio measurement levels (e.g., number of cigarettes smoked a week) sometimes exhibit right skewness in populations, in which case adding `dist = 'lnorm'` as a parameter may be warranted to provide a more accurate estimate of the required d given an MCD.

Further note that the relationship between MCD and Cohen's d is dependent on the CER value. For normally distributed outcome variables, it holds that the further the CER value is away from the mean of the distribution, a larger Cohen's d would be required to achieve an MCD of 5%. This is because in normal distributions, a CER of .50 implies a threshold value around the distribution mean. Therefore, the maximum number of people will be distributed around the threshold value – resulting in relatively small Cohen's d value for a given MCD. The further the CER deviates from the mean value (CER = .50), the larger Cohen's d must be to create a meaningful change. Figure 2 further illustrates this dependency between the MCD, the base-rate (CER) and the SESOI estimate in terms of Cohen's d (see also Gruijters & Peters, 2019). In case the population distribution of the outcome is for instance lognormal (right skewed), then a given MCD will require the smallest Cohen's d for CER values lower on x . Conversely, on the right end of the distribution tail, the largest d values are then required to generate an MCD.

STEP 5: Plan your sample size

Once the MCD-based Cohen's d is established, the required sample size can be computed using power or AIPE computations described in a previous section (see also

Table 3. Summary of the steps in the MCD-procedure.

Step #	Action
Step 1	Determine a threshold definition (TD) of positive events on the continuous outcome
Step 2	Ascertain or estimate the base-rate occurrence (CER) of positive events given the threshold definition.
Step 3	Set a meaningful change definition (MCD)
Step 4	Estimate Cohen's d based on the MCD using Equation 1
Step 5	Use the calculated Cohen's d as the SESOI to plan an appropriate sample size given power or small margin of errors

Cohen, 1988; Cumming, 2014; Faul et al., 2007; Peters and Crutzen, 2020). To illustrate the ball park, [Table 2](#) shows the required sample sizes for control event rates of 5%, 25% and 50%, and MCDs of 2.5%, 5%, 10%, and 25%, the required sample sizes to obtain 80% and 95% power and to estimate the intervention effect with 95% confidence interval half-widths of $d=.10$ and $d=.25$.

Discussion and recommendations

The MCD-procedure outlined in this paper enables applied researchers to determine the smallest effect of interest, for which the required sample size can then be planned (see [Table 3](#) for a summary of the procedure). The essence of the procedure is three-fold: 1) applied researchers need to have some *a priori* idea about what range of values on a continuous intervention outcome are considered 'positive' versus 'negative', 2) researchers need to have an estimate of the base-rate occurrence (CER) of these events, and 3) researchers need to set (or, obtain) a meaningful change definition – the increase in percentage of positive events considered meaningful change due to intervention. The MCD-procedure, besides giving intervention researchers a concrete method to arrive at a SESOI, highlights several practical considerations that are important when planning experimental tests of interventions. First, the procedure stresses the relevance of prevalence; in order to adequately test the efficacy of an intervention, information on the current state of the population improves the prediction of how a given effect size will affect this population.

The practical implication for normally distributed outcomes is this (illustrated in [Figure 2](#)): interventions aiming to change a behaviour with a very high or very low base-rate corresponding to the threshold definition will require larger effect sizes to engender meaningful change (see also Furukawa & Leucht, 2011; Gruijters & Peters, 2019). Thus, implying that interventions targeting outcomes with a sufficient (but not excessively small or large) base-rate in the population will tend to require relatively smaller effect sizes. This population-level thinking can be made somewhat more intuitive with an analogy: Suppose that we aim to improve jumping ability in a normally distributed population with a mean jumping skill of 1 meter 50 (sd = 20 cm). If we define a successful intervention as 'able to jump at least 1 meter 90', it will require a strong intervention to substantially increase the proportion of people able to do so – because few people will be anywhere close to this threshold. However, if we define intervention success as 'able to jump 1 meter 60' it will not require as strong an intervention to get a substantial number of people to jump more than 1 meter 60.

Requiring a smaller effect size is, of course, somewhat of a mixed blessing. On the one hand, from a practical perspective, this means that the intervention development

process is less demanding. For example, instead of requiring powerful behaviour change principles that need to be administered by an intermediary (e.g. a lifestyle coach), perhaps a mass media campaign can suffice. However, detecting small effect sizes using null hypothesis significance testing, or estimating small effect sizes with an accuracy commensurate to that small effect size magnitude, requires considerable sample sizes. If such sample sizes are not feasible, one alternative approach would be to develop a more efficacious intervention than is required to meet the established MCD. The somewhat prohibitive problem of this approach is the lack of theories that enable quantitative predictions of the effectiveness of behaviour change principles.

However, there is another – perhaps more practical – way to deal with small SESOI values requiring unfeasibly large samples. The MCD-method allows a concrete way to estimate the SESOI, but as implied by the term, this estimate is the lower-bound effect size of interest. This implies that finding robust evidence that the intervention effect size is larger than the SESOI (that is, $d > d_{MCD}$) warrants termination of the intervention efficacy test. For instance, given a SESOI of $d = 0.13$, the required sample size for a significance test with sufficient power – or a high precision estimate – is exceedingly large. One viable approach to deal with power analysis in case of a small SESOI is to use sequential analyses (for an accessible introduction, see Lakens, 2014). The basic idea behind sequential analyses (see also Albers, 2019; Neumann et al., 2017) is as follows. Consider a project in which the estimated sample size for a sufficiently powered test amounts to $n = 375$. Going the whole nine yards in data collection may (depending on the specific design) involve considerable resources (e.g., time and money). Researchers could instead plan for interim analyses at various pre-registered intervals during data collection (e.g., after $n = 125$, $n = 250$). When following proper protocol and controlling for the false positive rate (see Albers, 2019; Lakens, 2014), interim analyses allow researchers to terminate data collection when sufficient evidence for intervention effectiveness, or absence thereof, has been found (e.g., Albers, 2019; Lakens, 2014; Neumann et al., 2017).

In conclusion, the MCD-method further illustrates the importance of undergirding the intervention research with concrete and specific goals: what outcomes of the intervention will be deemed ‘positive’ and on what criteria are such choices based? This, in turn requires interventionists to consider the real-world impact they want to achieve. Finally, the MCD-method emphasizes the importance for intervention researchers to involve stakeholders and policymakers external to the project in order to set such goals.

Acknowledgements

The authors are thankful to two anonymous reviewers for constructive comments on a previous version of this manuscript.

Author contributions

S.L.K. Gruijters generated the idea for the MCD-method. G-J.Y. Peters wrote the R-code and scripts. S.L.K. Gruijters wrote the first draft of the manuscript, G-J.Y. Peters wrote parts of the

manuscript, and both authors critically edited the entire manuscript. Both authors approved the final submitted version of the manuscript

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Stefan L. K. Gruijters  <http://orcid.org/0000-0003-0141-0071>

Gjalt-Jorn Y. Peters  <http://orcid.org/0000-0002-0336-9589>

References

- Aguinis, H., & Harden, E. (2009). Sample size rules of thumb: Evaluating three common practices. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 267–286). Taylor & Francis.
- Albers, C. (2019). The problem with unadjusted multiple and sequential statistical testing. *Nature Communications*, 10(1), 1921. <https://doi.org/10.1038/s41467-019-09941-0>
- Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28(11), 1547–1562. <https://doi.org/10.1177/0956797617723724>
- Anderson, S. F., & Maxwell, S. E. (2017). Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3), 305–324. <https://doi.org/10.1080/00273171.2017.1289361>
- Anvari, F., & Lakens, D. (2019). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. <https://doi.org/10.31234/osf.io/syp5a>
- Bartholomew Eldregde, L. K., Markham, C. M., Ruiter, R. A. C., Fernández, M. E., Kok, G., Parcel, G. S. (2016). *Planning health promotion programs: An Intervention Mapping approach* (4th ed.). Wiley.
- CBS/RIVM. (2018). *Leefstijl monitor*. <https://www.volksgezondheidenzorg.info/onderwerp/sport-en-bewegen/cijfers-context/huidige-situatie#node-beweegrichtlijnen>
- Centraal Bureau voor de Statistiek. (2019). *StatLine - Bevolking; generatie, geslacht, leeftijd en migratieachtergrond, 1 januari*. Retrieved July 24, 2019, from <https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37325/table?ts=1563955220702>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3), 145–153. <https://doi.org/10.1037/h0045186>
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7(3), 249–253. <https://doi.org/10.1177/014662168300700301>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Quantitative Methods in Psychology*, 112(1), 155–159.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
- Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. Lance & R. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 287–308). Taylor & Francis. <https://doi.org/10.4324/9780203867266>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>

- DeCoster, J., Iselin, A.-M R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods, 14*(4), 349–366. <https://doi.org/10.1037/a0016956>
- Driessen, E., Hollon, S. D., Bockting, C. L. H., Cuijpers, P., & Turner, E. H. (2015). Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-funded trials. *PLOS One, 10*(9), e0137864. <https://doi.org/10.1371/journal.pone.0137864>
- Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research, 51*(5), 589–605. <https://doi.org/10.1080/00273171.2016.1191324>
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891–904. <https://doi.org/10.1007/s11192-011-0494-7>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. <https://doi.org/10.3758/bf03193146>
- Furukawa, T. A., & Leucht, S. (2011). How to obtain NNT from Cohen's d: Comparison of two methods. *PLoS One, 6*(4), e19070. <https://doi.org/10.1371/journal.pone.0019070>
- Gezondheidsraad. (2017). *Beweegrichtlijnen 2017*. <https://www.gezondheidsraad.nl/documenten/adviezen/2017/08/22/beweegrichtlijnen-2017>
- Grujters, S. L. K., & Peters, G.-J Y. (2019). *Gauging the effectiveness of behavior change interventions: A tutorial on the number needed to treat*. <https://doi.org/10.31234/osf.io/2bau7>
- Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., & Norman, G. R. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77*(4), 371–383. <https://doi.org/10.4065/77.4.371>
- Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. *Controlled Clinical Trials, 10*(4), 407–415. [https://doi.org/10.1016/0197-2456\(89\)90005-6](https://doi.org/10.1016/0197-2456(89)90005-6)
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*(4), 363–385. <https://doi.org/10.1037/1082-989X.11.4.363>
- Kelley, K. (2020). *MBESS: The MBESS R package*. R package version 4.8.0. <https://cran.r-project.org/package=MBESS>
- Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry, 63*(5), 484–489. <https://doi.org/10.1001/archpsyc.63.5.484>
- Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *The British Journal of Mathematical and Statistical Psychology, 65*(2), 350–370. <https://doi.org/10.1111/j.2044-8317.2011.02029.x>
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*(7), 701–710. <https://doi.org/10.1002/ejsp.2023>
- Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability. *Perspectives on Psychological Science, 9*(3), 278–292. <https://doi.org/10.1177/1745691614528520>
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. <https://doi.org/10.1037/1082-989X.7.1.19>
- Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331–348. <https://doi.org/10.2466/03.11.PMS.112.2.331-348>
- McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: Defining what really matters to patients. *JAMA, 312*(13), 1342–1343. <https://doi.org/10.1001/jama.2014.13128>
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. <https://doi.org/10.1086/288135>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*(3), 221–229. <https://doi.org/10.1038/s41562-018-0522-1>

- Neumann, K., Grittner, U., Piper, S. K., Rex, A., Florez-Vargas, O., Karystianis, G., Schneider, A., Wellwood, I., Siegerink, B., Ioannidis, J. P. A., Kimmelman, J., & Dirnagl, U. (2017). Increasing efficiency of preclinical research by group sequential designs. *PLoS Biology*, *15*(3), e2001307 <https://doi.org/10.1371/journal.pbio.2001307>
- Peters, G.-J. Y., & Crutzen, R. (2020). Knowing how effective an intervention, treatment, or manipulation is and increasing replication rates: accuracy in parameter estimation as a partial solution to the replication crisis. *Psychology & Health*, 1–19. <https://doi.org/10.1080/08870446.2020.1757098>
- Peters, G.-J. Y., Crutzen, R. & Gruijters, S.L.K. (2020a). *Behaviorchange: Tools for behavior change researchers and professionals*. R package version 0.2.4. <https://cran.r-project.org/package=behaviorchange>
- Peters, G.-J. Y. & Gruijters, S.L.K. (2020b). *Ufs: Quantitative analysis made accessible*. R package version 0.3.2. Retrieved from <https://cran.r-project.org/package=ufs>
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*(2), 166–169. <https://doi.org/10.1037/0022-0663.74.2.166>
- Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*(6), 775–777. <https://doi.org/10.1037/0003-066X.45.6.775>
- Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist*, *46*(10), 1086–1087. <https://doi.org/10.1037//0003-066X.46.10.1086>
- Smaldino, P. E. (2017). Models Are Stupid, and We Need More of Them. In R. Vallacher, S. Read, & A. Nowak (Eds.), *Computational Social Psychology* (pp. 311–331). New York : Routledge, 2017. | Series: *Frontiers of social psychology*: Routledge. <https://doi.org/10.4324/9781315173726-14>
- U.S. Department of Health and Human Services. (2018). *Physical activity guidelines for Americans title*. https://health.gov/paguidelines/second-edition/pdf/Physical_Activity_Guidelines_2nd_edition.pdf
- Viechtbauer, W., Smits, L., Kotz, D., Budé, L., Spigt, M., Serroyen, J., & Crutzen, R. (2015). A simple formula for the calculation of sample size in pilot studies. *Journal of Clinical Epidemiology*, *68*(11), 1375–1379. <https://doi.org/10.1016/j.jclinepi.2015.04.014>