The Journal of Child Psychology and Psychiatry

Journal of Child Psychology and Psychiatry **:* (2023), pp **–** doi:10.1111/jcpp.13817

Editorial Perspective: When is a 'small effect' actually large and impactful?

Emma Grace Carey,[1] Isobel Ridler,[2] Tamsin Jane Ford,[1] and Argyris Stringaris[2,3]

[1]Department of Psychiatry, University of Cambridge, Cambridge, UK; [2]Division of Psychiatry, University College London, London, UK; [3]First Dept of Psychiatry, National and Kapodistrian University of Athens, Greece

Introducing effect sizes and their use in mental health research

The past three decades have seen a dramatic shift towards reporting effect sizes, such as Cohen's *d*, that convey information about the magnitude of the relationship between variables (Schäfer & Schwarz, 2019). In the case of the pandemic, clinicians, policy makers and the public want to know what effects events such as school closures have had on youth mental health (Ford, John, & Gunnell, 2021; Mansfield et al., 2022). This leads to the issue of how to evaluate effect sizes: in the case of the pandemic for example, how to interpret the magnitude of change in mental health problems over time in relation to different phases of the pandemic. In this article, we review some of the issues with the reporting and interpretation of effect sizes and present some simulations to illustrate common problems.

Problems with interpreting effect sizes

Despite the obvious relevance and importance of effect sizes to psychological research, and the additional information conveyed by reporting these alongside measures of statistical significance, some standard interpretations of effect sizes can be misleading if used in the wrong context. Traditionally, effect sizes have been reported in two ways, both of which can be problematic to interpret. Cohen's *d* interprets mean differences between two time points or two groups by scaling them with the standard deviation, a measure of the variability of the data. The bigger the mean differences (or the smaller the standard deviations), the bigger the effect size. Jacob Cohen lived to regret (R. Rosenthal, personal communication, November 2018) his standard conventions, which determined *r* values of 0.10, 0.30 and 0.50 and *d* values of 0.2, 0.5 and 0.8 as representing *small*, *medium* and *large* effects, respectively (Cohen, 1977, 1988, 1992). Whilst these were used by Cohen, apparently reluctantly (Funder & Ozer, 2019) in the context of power analyses where no better option was available, they have become ubiquitous 'rules' in psychology research (Funder & Ozer, 2019). This is problematic since the terms small, medium and large are meaningless without a frame of reference or comparison (see Textbox 1 for examples of different contexts).

To illustrate this, we will take the small effect size of $d = 0.14$ found by Mansfield et al. (2022), as our example to explain frames of reference. The size of the *d* is also relevant to the pandemic-related discussion that follows.

For a start, let us translate this small effect size into actual Moods and Feelings Questionnaire (MFQ) points. Given the population mean of the MFQ is $mean\_1 = 4.92$ and its standard deviation $SD = 4.49$ (Kwong, 2019), an effect size of around $d = 0.14$ would mean a shift to a $mean\_2 = 5.55$ (at the same *SD*). The difference between MFQs would be 0.63 points. An effect size of $d = 0.22$, which is still considered a small effect, would lead to a difference of 1 MFQ points. We have included additional simulations at this effect size to highlight the population level effect of a shift of just 1 point on the MFQ – these can be found in the Table S1.

Small but meaningful effect sizes in the pandemic

Mental health services in the United Kingdom have long been stretched (Fonagy & Pugh, 2017). Since the start of the COVID-19 pandemic and subsequent lockdowns, social isolation, school closures and loss of health and lives, the declining mental health of children and young people seems to be pushing these services to breaking point. Presentations of young people to Child and Adolescent Mental Health Services (CAMHS) have increased, particularly in relation to eating disorder (Ford et al., 2021). As it were, the magnitude of the impact of this deterioration in mental health on the presentation of young people and families seems big, with a doubling, for example, in eating disorder referrals between 2020 and 2021 (Solmi, Downs, & Nicholls, 2021). However, in contrast, the effect sizes reported on measures of population mental health seem rather small. Some studies reassuringly describe negative effects as 'small' based on Cohen's effect size judgement. For instance, examining the difference in depression scores before and during the pandemic found an effect size of $d = 0.14$ (Mansfield et al., 2022). In light of these findings, some will wonder why services are struggling with increased pressure. There are many explanations for this increased pressure on services –

Conflict of interest statement: No conflicts declared.

Textbox 1: Effect Sizes in Different Contexts

Effect sizes in a clinical context

- Clinicians care and make decisions about individual patients.
- They are encouraged to use outcome measures, typically summed scores on questionnaires, such as the Mood and Feelings Questionnaire (MFQ), to assess patients' progress in treatment.
- To any clinician, a change in a total score of around one point on the MFQ would seem negligible and would not be a basis to act. Similarly, no clinician would deem patient A scoring 15 points as being significantly more unwell compared to a patient B who scores 14 points.
- In this instance, a "small" effect size is genuinely small and is not taken to represent a meaningful or significant shift.

Effect sizes in a clinical trial

- Even clinical trialists, who may typically deal with 100s of participants, would think very little of an MFQ difference of 1 point in the MFQ or a $d = 0.22$ (an effect size that would correspond to it see below).
- A novel pharmacological compound in a typical depression trial producing such an effect size would probably be deemed a failure.
- Moreover, even if there were a true difference between a group that receives the new compound and the group receiving placebo, this difference would require a very large sample size to be demonstrated (by a simple power calculation for typical power required about $n = 800$ per group, very unusual for most trials).
- Understandably therefore, neither a clinician nor a trialist would make much of such a change under usual circumstances and Cohen's rules of thumb about effect sizes would apply. The problem is, however, that this mode of thinking is then extended to the public health domain.

Effect sizes in research with an entire population

- A public health specialist will need to view effect sizes differently. What may seem small to a clinician and a trialist, could have profound effects if it occurs in the general population.
- The reason is relatively simple, even though often misunderstood, and is best illustrated when thinking about how many people would cross a threshold for any given mean shift in a population.
- Let us assume a threshold of the MFQ above which we declare someone as being depressed. The *relative* number of people who cross that threshold given an effect size is the same across any population size? however, the *absolute* number will vary dramatically by population size. We illustrate this below using the effects of the pandemic as an example.

for example, there may have been a gradual rise in unmet need, change in demographics of the population, increased referrals of young people with mental health difficulties, or lower capacity of services. In this article, we consider one possible explanation for the increased demand for services and resultant pressure on CAMHS: that a small shift in mean scores may disproportionately affect the tail of the distribution. For CAMHS, this will mean disproportionately more children and young people with poor mental health will now exceed CAMHS thresholds.

Alternative metrics do exist for the evaluation of change in a population distribution. For example, Rom and Hwang (1996) discuss the use of the proportion of similar responses (PSR) to evaluate the change between two distributions (pre-treatment vs. post-treatment). The PSR provides a measure of the overlap between two probability distributions, with a PSR of 0 corresponding to nonoverlapping distributions and a PSR of 1 corresponding to two perfectly overlapping distributions. Whilst this has typically been used to evaluate normal distributions, nonparametric estimates also exist (Giacoletti & Heyse, 2011). In this paper, we use simulations to evaluate the potential population-based effects of 'small' effect sizes, but these alternative metrics may also be useful.

Simulating pandemic data

To address the question of how a small shift in means could affect the tail of the distribution on psychological measures, and therefore have a disproportionate effect on the number of cases presented to CAMHS, we have utilised simulation of data from the short form MFQ (MFQs). The MFQs had a pre-pandemic mean of 4.92 ($SD = 4.49$; Kwong, 2019). For an effect size of around $d = 0.14$, the post-pandemic mean would be 5.55 (retaining $SD = 4.49$; the effects described below are exacerbated if the standard deviation is also allowed to increase). The standard threshold which indicates likely presence of depression is a score of 2

Table 1 Excess cases by population size affected of mood disorders in young people from before to during the Covid-19 pandemic according to simulated Moods and Feelings Questionnaires data. Estimates from simulations with 1,000 repetitions

Number affected	Pre-pandemic cases (*SD*)	Pandemic cases (*SD*)	Change in cases (*SD*)
100	10.14 (2.92)	11.73 (3.23)	1.58 (4.41)
1,000	101.32 (9.32)	117.58 (10.57)	16.26 (13.86)
10,000	1,012.78 (30.57)	1,174.18 (32.70)	161.40 (45.61)
100,000	10,123.47 (94)	11,732.08 (98.89)	1,608.61 (132.43)
1,000,000	101,194.23 (291.34)	117,279.67 (324.88)	16,085.44 (438.73)
10,000,000	1,012,035.52 (947.47)	1,172,905.21 (1,002.62)	160,869.70 (1,348.08)

or more on the MFQs. We simulated data to reflect both pre-pandemic and post-pandemic MFQs means, using a scaled beta distribution to reflect the typical right skew and bounded nature of the MFQs (Kwong, 2019; only scores between 0 and 26 are possible; note alternative distributions, such as the normal lead to similar or exacerbated effects as the ones we describe below). We will use our simulated data to quantify how many additional cases of depression in young people would be expected to occur at the given effect size, an approach which has previously been dissussed in Grice et al. (2020).

Table 1, row 5, shows the simulated results for pre-pandemic and post-pandemic scores during a pandemic affecting 1 million young people and shows that whilst the mean MFQs score has only changed a small amount (4.92–5.55), there is a disproportionate increase in the tail of high scores (from 10.1% of responses to 11.7% of responses). This reflects an increase of 160,870 cases of depression at a population size of 10 million young people.

Of course, there are many more than 1 million children and young people in the UK. In 2019, there were 12.7 million people under the age of 16 living in the UK. We have therefore scaled up the simulation results, to see the implication of a small shift in mean scores on caseload across the UK. Table 1 shows the number of extra cases of depression that could be expected based on the number of children impacted by the pandemic. With more than 10 million young people impacted by the pandemic in the UK, based on our simulations we could expect around than 160,000 extra cases of depression which may present to CAMHS or other services. In reality, because more adolescents experience symptoms of depression than young children (Hoare et al., 2020), this could be an overestimate – however, we would still expect an excess in CAMHS referrals. With 397,822 referrals to CAMHS in 2019–20 where we would expect a prevalence of 1 million young people having depression based on simulations, it appears that around 40% of those meeting the MFQ threshold end up being referred to CAMHS. An additional 160,000 cases of depression, with a continued 40% referral rate, would represent an increase in referrals of 64,000. This would represent a 16% increase in CAMHS referrals (perhaps more if the referral rate has also increased) and may explain why the service is breaking under this increased pressure. Real-life data show that with 527,339 referrals to CAMHS between 2020 and 2021, referrals were increased by 33% compared with the previous year. Our findings suggest that this increase in referrals is likely to be driven in considerable part by a relatively small shift in average MFQ scores.

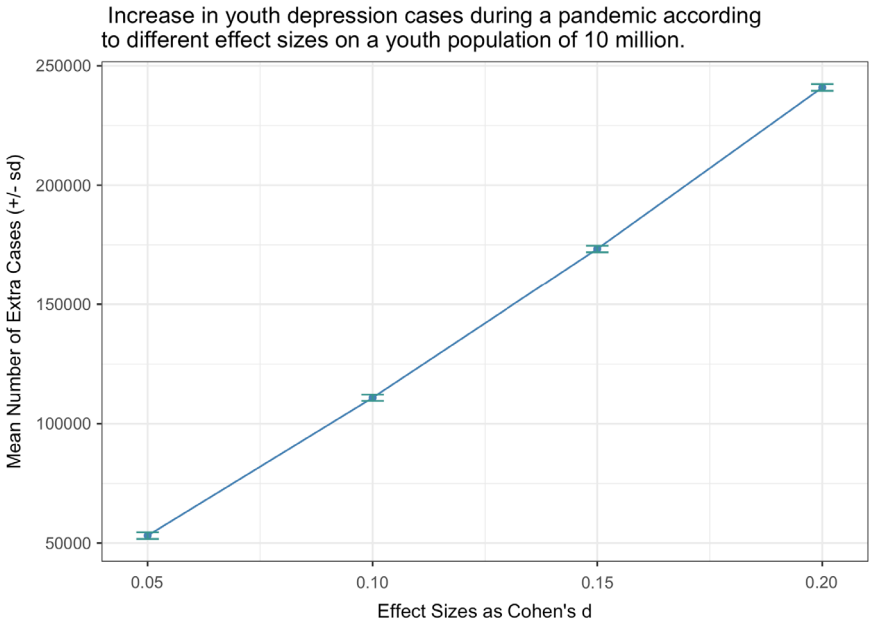

Figure 1 Excess depression cases during a pandemic by number of youth affected and effect size according to simulated Moods and Feelings Questionnaires data. Estimates from simulations with 1,000 repetitions

We ran further simulations to demonstrate how shifts in mental health instrument results with effect sizes of between 0.05 (considered very small) and 0.2 would affect the number of extra cases of depression in young people. Figure 1 displays the number of extra cases of depression expected depending on the number of young people affected by the pandemic for six different levels of effect size. This graph demonstrates that even for a very small effect size of 0.05, with 10 million young people affected by the pandemic, one would expect more than 50,000 extra cases of depression – not an insignificant number for a service which is already stretched. All effect sizes on this graph would be labelled as small, but it is clear to see that their material impact may be very substantial.

Other authors have made similar arguments that small effect sizes when applied to an entire population may scale up to be impactful. For example, Götz et al. (2022) argued that small effect sizes are the norm in psychological sciences and can be highly relevant. Primbs et al. (2022) make the important point that this should not be used as an argument to uncritically accept all and any small effect as important or impactful, which could be dangerous. Similarly, Anvari et al. (2022) argue that it is important to consider that when an effect size is generalised to a new context, one must consider both amplifying *and counteracting* mechanisms rather than heuristically accepting that all effects are important. These arguments emphasise the importance of nuance and caution when interpreting small effect sizes: empirical evidence and/or a falsifiable line of theoretical reasoning should be present when making the argument that a small effect is important or impactful, in order to avoid a proliferation of non-replicable findings.

Conclusions

Others have made the point that classifying effect sizes as small, medium and large is meaningless without a frame of reference. We have taken this further to explain the initially confusing finding that use of services by children and young people with mental health conditions in the UK has dramatically increased during the pandemic, despite a relatively small effect size on measures of psychopathology. Particularly in the field of public health, small effects scaled across an entire population can be very relevant and impactful when supported by empirical evidence or strong, falsifiable lines of reasoning. Furthermore, a modest shift in effect sizes can have a disproportionate effect on the tail of the distribution. In the case of psychiatric classification where a 'case' is someone with an extremely high score, this could mean that a small shift in mean score has a profound effect on case numbers that seems likely to feed into additional referrals. Our work also demonstrates the value of simulation in quantifying the genuine effect portrayed by a 'small' effect size. This "small effect big impact" phenomenon also has a positive side when considering interventions that are similarly scalable: universal intervention with small effects may also be expected to have big salutary effect (Greenberg & Abenavoli, 2017).

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Table S1. Excess cases by population size affected of mood disorders in young people from before to during the COVID-19 pandemic according to simulated MFQs data.

Acknowledgements

The authors have declared that they have no competing or potential conflicts of interest. We would like to thank Max Primbs for making us aware of literature which highlights the importance of critically analysing small effect sizes and additional research which uses the approach of quantifying effect sizes in terms of their person-based effects.

Correspondence

Emma Grace Carey, Department of Psychiatry, University of Cambridge, Clifford Allbutt Building, Addenbrookes, Hills Road, Cambridge CB2 0XY, UK; Email: ec475@cam.ac.uk

Key points

- Effect sizes reported in psychology and psychiatry research are often interpreted according to standard benchmarks for 'small', 'medium' and 'large' effects.
- In reality, a 'small' effect size can have large and meaningful impacts, particularly when applied to large populations.
- We show using simulations that a 'small' effect size relating to change in MFQ score for an entire population can result in many excess cases of mental ill health.
- This explains why a 'small' change in mean MFQ scores from before to during the COVID-19 pandemic could result in mental health services becoming overwhelmed with excess cases.
- This research both demonstrates the need for more nuanced contextual understanding of effect sizes and the use of simulations in this context.

References

Anvari, F., Kievit, R., Lakens, D., Pennington, C.R., Przybylski, A.K., Tiokhin, L., Wiernik, B.M., & Orben, A. (2022). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916221091565

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences, Rev. ed.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Fonagy, P., & Pugh, K. (2017). Editorial: CAMHS goes mainstream. *Child and Adolescent Mental Health*, *22*, 1–3.

Ford, T., John, A., & Gunnell, D. (2021). Mental health of children and young people during pandemic. *BMJ*, *372*, n614.

Funder, D.C., & Ozer, D.J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156–168.

Giacoletti, K.E.D., & Heyse, J. (2011). Using proportion of similar response to evaluate correlates of protection for vaccine efficacy. *Statistical Methods in Medical Research*, *24*, 273–286.

Götz, F.M., Gosling, S.D., & Rentfrow, P.J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*, 205–215. https://doi.org/10.1177/1745691620984483

Greenberg, M.T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, *10*, 40–67.

Grice, J.W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M., (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, *3* (4), 443–455. https://doi.org/10.1177/2515245920922982.

Hoare, E., Werneck, A.O., Stubbs, B., Firth, J., Collins, S., Corder, K., & Van Sluijs, E.M.F. (2020). Association of Child and Adolescent Mental Health with Adolescent Health Behaviors in the UK millennium cohort. *JAMA Network Open*, *3*, e2011381.

Kwong, A.S.F. (2019). Examining the longitudinal nature of depressive symptoms in the Avon longitudinal study of parents and children (ALSPAC). *Wellcome Open Research*, *4*, 126.

Mansfield, R., Santos, J., Deighton, J., Hayes, D., Velikonja, T., Boehnke, J.R., & Patalay, P. (2022). The impact of the COVID-19 pandemic on adolescent mental health: A natural experiment. *Royal Society Open Science*, *9*, 211114.

Primbs, M.A., Pennington, C.R., Lakens, D., Silan, M.A.A., Lieck, D.S.N., Forscher, P.S., Buchanan, E.M., & Westwood, S.J. (2022). Are small effects the indispensable foundation for a cumulative psychological science? A reply to Götz et al. (2022). *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916221100420

Rom, D.M., & Hwang, E. (1996). Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine*, *15*, 1489–1505.

Schäfer, T., & Schwarz, M.A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813.

Solmi, F., Downs, J.L., & Nicholls, D.E. (2021). COVID-19 and eating disorders in young people. *The Lancet Child & Adolescent Health*, *5*, 316–318.

Accepted for publication: 19 March 2023