Time to Overcome the Neglect of Effect Sizes in Teaching Psychological Research Findings

Johannes Hönekopp[*1], and Joanna Greer[1]

Scholarship of Teaching and Learning in Psychology, in press.

[1] Department of Psychology, Northumbria University.

[*]Corresponding author information:

Johannes Hönekopp

Northumbria University

Department of Psychology

Newcastle upon Tyne, NE1 8ST

United Kingdom

johannes.honekopp@unn.ac.uk

++44 191 243 7478

Acknowledgements: We would like to thank Frank Renkewitz and Tamsin Saxton for helpful comments on an earlier version of this paper.

Time to Overcome the Neglect of Effect Sizes in Teaching Psychological Research Findings

Abstract: Due to reliance on null-hypothesis significance testing, researchers had widely neglected effect size (ES) in their reporting of psychological research results. The American Psychological Association and others identified this as a problem and initiated reform, and ES reporting is now common in the scholarly literature. To what extent problematic ES neglect occurs in the teaching of psychological research findings remains unknown. In order to address this question, we analysed the content of 10 bestselling psychology textbooks from five popular sub-fields (Study 1). Overall, 76% (95% CI [69%, 79%]) of portrayals of empirical findings were devoid of useful ES information and merely mentioned the direction of the effect discussed. Such a direction-only description of any research finding would appear particularly uninformative if laypeople's intuitions about the direction of this effect were already accurate. Study 2 therefore tested how well 286 non-psychology undergraduate students could guess the direction (or absence) of a representative sample of effects. Irrespective of psychological sub-field of the effects, average accuracy was only 47% [42%, 51%], demonstrating participants' poor intuitions regarding these effects. This suggests that the predominant direction-only description of research findings in textbooks is not without merits. However, we argue that greater consideration of ES in the teaching of psychological research findings is desirable. We suggest three simple and efficient strategies to this end and weight their relative merits.

Key words: effect size; critical thinking; statistical reform; teaching; psychology.

The reporting versus neglect of the magnitude of effects in psychological research has received considerable attention (Fritz, Scherndl, & Kühberger, 2013). Curiously, to what extent the magnitude of effects is communicated in psychology teaching has been ignored. Effect sizes (ESs) are widely covered in statistics textbooks (Capraro & Capraro, 2002), which suggests that undergraduates will typically be familiar with this concept. But it is currently unknown to what extent the magnitude of observed effects is addressed in psychology teaching beyond methods training: For example, when students encounter Asch's (1956) famous conformity experiment in their social psychology class, will they typically learn *how much* participants' perceptual judgements suffered as the result of majority pressure? Or will they merely learn that there *was* such an effect? Here, we argue that psychology teaching should address the strength of the effects being taught. In Study 1, we explore for the first time to what extent teachers convey ESs outside method classes, and how they achieve this. We then argue that the relative merit of different communication strategies hinges in part on learners' prior knowledge of the effects covered. To address this, in Study 2, we explore laypeople's knowledge about findings taught in psychology. We then use these findings to discuss the usefulness of various strategies for communicating the magnitude of the effects taught in psychology .

Null-hypothesis significance testing (NHST) has dominated data analysis in psychology since the 1950's (Fritz, Morris, & Richler, 2012). NHST seeks to show that an observed (or a more extreme) sample effect (a difference between group means, a correlation, etc.) would be unlikely if the true population effect was zero. For a long time, critics have stressed that overreliance on NHST in the analysis and interpretation of data invites trouble. For example, a focus on *p*-values makes meaningful comparisons of results across studies next to impossible (Hunter, 1997; Schmidt, 1992). Also, NHST uses the data to challenge the null-hypothesis, which – in psychology – rarely reflects the researcher's hypothesis. The

latter then remains unchallenged by the data, and little insight is gained when the (statistically significant) test rejects what seemed implausible in the first place (Meehl, 1967). Moreover, NHST misleads researchers to neglect the strength of effects (Tukey, 1991; Yates, 1951), which hinders both theoretical progress and judgements of practical usefulness (e.g. Cohen, 1994; Ferguson, 2009; Kirk, 1996; Meehl, 1978).

In response to criticism of NHST, the *APA* pushed for reform and, among the suggested remedies, consideration of ESs takes a prominent place (Appelbaum et al., 2018; Wilkinson & Task Force on Statistical Inference, 1999). The most recent edition of the *APA Publication Manual* (American Psychological Association, 2010) stresses that:

"NHST is but a starting point [for data analysis] … For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size" (pp. 33-34).

An overview of 29 surveys, covering the years 1990 to 2007, found steady increases (~2% per year) in standardised ES reporting across a broad selection of psychology journals, with most papers reporting them from 2004 onwards (Fritz et al., 2013). Thus, presumably in response to reform efforts, researchers increasingly go beyond NHST's question '*Is there an effect?*', and provide ESs to describe its magnitude.

It seems obvious that ES also matters when students are taught about central findings in psychology. We refer again to Asch's famous conformity experiment (1956), where participants provided blatantly wrong perceptual judgements in response to majority pressure. Importantly, these errors occurred *frequently* (experimental group 37% vs. control group 5%), rather than as the occasional – though statistically significant – exception.

ESs also have immediate practical implications. Consider a recent large-scale study into the wellbeing of 15-year olds in the UK as an example (Gireesh, Das, & Viner, 2018). Substance use, unhealthy eating habits, screen time, physical activity, and other modifiable behaviors were investigated, and all of them turned out to be statistically significant predictors of wellbeing. However, getting enough sleep proved to be considerably more important than any other predictor. And because sleep is a *particularly strong* predictor of wellbeing, it suggests itself as an especially promising target for interventions.

These two examples illustrate why ES is a central aspect of any study. Accordingly, we think that the strength of psychological findings should be communicated not only in the scholarly literature but also when they are taught to students. To what extent does this happen? We address this issue in Study 1; specifically, we look at psychology textbooks. Due to their large readership and wider impact (informing lectures and other modes of teaching), textbooks are particularly important. We focussed on best-selling textbooks in biological, cognitive, developmental, personality, and social psychology, because these fields enjoy great popularity among undergraduates (Norcross et al., 2016). We share the view that open science supports the advance of psychological knowledge (e.g. Shrout & Rodgers, 2018); therefore, all materials and data can be accessed at https://osf.io/uv4f3/.

Study 1: Effect Size Communication in Textbooks

Method

Selection of textbooks. We searched www.amazon.com (6th January 2015) to source the two bestselling student textbooks each for biological (Gazzaniga, Ivry, & Magnum, 2013; Kalat, 2012), cognitive (Goldstein, 2014; Sternberg & Sternberg, 2012), developmental (Berger, 2014; Santrock, 2013), personality (Feist, Feist, & Roberts, 2012; Larsen, Buss, & Wismeijer, 2013), and social psychology (Aronson & Wilson, 2012; Myers, 2012).

Sampling of textbook passages. To avoid clustering from specific sections, we arbitrarily selected from each textbook 40 pages that spanned the entire book and were separated by similar intervals. From each page, we selected the first topic that met our inclusion criteria (see Table 1) for analysis. Typically, textbooks will dedicate more space to particularly important research findings. Our page-based sampling approach therefore ensured that central findings were more likely to enter our sample than peripheral ones. Overall, 134 pages did not contain relevant content and 266 textbook passages entered our analysis. All selected passages within the same textbook dealt with different research findings.

Once a topic was selected, we considered all relevant text on the selected page (including overhanging paragraphs from the previous or following page) and pertinent figures on the selected, the previous, and the following page,. Amongst the relevant text and/or figure, we coded the part most informative about ES for analysis (see below).

Coding of textbook passages. The magnitude of an observed effect can be communicated by other means than a standardized ES, e.g. verbally ('sleep duration was a particularly strong predictor of wellbeing') or in original units ('participants in the treatment group lost, on average, 5.0 kg in weight'). We decided on a tripartite categorization, where both authors independently coded all passages as providing either *no*, *medium*, or *high* ES information. *No* passages stated at best the direction of the effect (e.g. "students with a mimicking rather than a nonmimicking digital companion […] liked the partner more", Myers, 2012, p. 235). *Medium* passages provided a vague verbal ES description (e.g. "the left hemisphere falsely recognized the new pictures related to the story, while the right hemisphere *rarely* [our emphasis] made that mistake", Gazzaniga, Ivry, & Magnum, 2013, p. 143), or provided an ES in raw score units that are difficult to understand for laypeople (e.g. "a two-point rise in students' Narcissistic Personality Inventory scores over two decades",

Aronson & Wilson, 2012, p. 131). Finally, *high* applied to passages that described ES in raw score units readily understood by laypeople (e.g. "the choice reaction time took one-tenth of a second longer than simple reaction time", Goldstein, 2014, p. 6), as proportions (e.g. % errors in Asch's 1956 conformity experiments), by means of a standardised ES measure (e.g. a correlation), or by any other means that provided an accurate understanding of the magnitude of the effect (e.g. "A survey […] showed that the risk of a collision was four times higher when the driver was using a cell phone than when a cell phone was not being used", Goldstein, 2014, p. 103).

For two out of 266 passages, a protocol error precluded independent coding. For the remaining, we computed agreement as weighted kappa, κ_w, using quadratic weights (Cohen, 1968). Computations were performed via an online calculator (www.vassarstats.net/kappa.html) and resulted in $\kappa_w = 0.73$ (SE = 0.04), which is typically interpreted as good (Jakobsson & Westergren, 2005). We resolved initial coding disagreements by discussion.

Results and Discussion

Figure 1 shows relative frequencies for the three levels of ES information across the 10 textbooks. Overall, *no* was most frequent (M = 47%[1], 95% CI [41%, 53]), followed by *medium* (M = 28%, [23%, 34%]) and *high* (weighted M = 26%, [21%, 33%]). Indeed, *no* was the modal category for 8 out of 10 textbooks. However, an exploratory analysis showed that results varied statistically significantly between authors (Kruskal-Wallis test, $\chi^2(9) = 21.5$, $p = .010$), and *medium* and *high* were the modal categories for one biological and one cognitive textbook, respectively. Large differences within the same subject area (cf. Figure 1) suggest

[1] We report weighted means throughout. Unweighted means were very similar.

that the observed differences between textbooks do not just reflect differences between disciplines, but also important differences in their authors' approaches.

To understand authors' strategies for conveying *high* levels of ES information, we analysed the relevant 68 passages in greater detail. Typically (57% [45%, 68%]), authors used percentages or proportions, similar to our portrayal of Asch's (1956) conformity experiment above; in another 24% [15%, 35%], authors reported means in informative raw score units. Standardised ESs, which are often seen as essential for original research reports, were virtually absent.

At the time of writing, eight of the textbooks in our sample had appeared in a new edition. In these later editions, we checked the authors' comments describing the changes they had made to understand how these new editions differed from the copies used in Study 1. We found no indication that ES reporting had changed. We therefore think it likely that our results still hold for later editions.

Our results are in stark contrast to reporting in psychology journals (Fritz et al., 2013), where a majority of papers now report standardised ESs (and presumably many more report unstandardized ESs). Excluding research methods, the apparent lack of ES information in psychology teaching seems regrettable, given that students should greatly benefit from ESs (see below).

An obvious limitation of our study is that our textbook sample is unrepresentative for psychology teaching in general. We can only speculate about the wider picture. On the one hand, textbooks are often recommended by expert teaching staff. Likely, their perceptions of textbook quality inform these recommendations; therefore, bestseller status should, to some extent, reflect favourable peer evaluation. It would then appear unlikely that other forms of psychology teaching (lectures, textbooks selling fewer copies, etc.) are systematically more rigorous about ESs. On the other hand, bestselling textbooks tend to be well-established (the

median textbook in our sample was in its 7th edition). Consequently, many of them might be first written when the importance of ES reporting was less discussed, and possibly their authors did not pay much attention to this issue in new editions. Our findings from Study 1 provide a starting point but this clearly needs greater empirical investigation.

We already argued that the magnitude of any psychological effect is of considerable importance, and should therefore be addressed in its teaching. From this viewpoint, the frequent neglect of ES in textbooks observed in Study 1 seems disappointing. However, we believe that in order to fully grasp the implications of this finding, it is necessary to take learners' prior knowledge about the taught phenomenona into account. To illustrate this point, we re-visit Gireesh et al.'s (2018) finding that sleep is a powerful predictor of wellbeing in British adolescents. Say we decided to teach this finding and we followed the modal practice, i.e. do so without any recourse to ES (e.g. 'The authors observed that adolescents who slept more had higher life satisfaction than those who slept less'). How much students can learn from our description now crucially depends on their assumptions about the sleep-wellbeing link. A student who thinks that the two are unrelated can learn something new from our description of the research. But it is difficult to see what our direction-only description of this finding could add to the knowledge of a student who already presumes that sleep and wellbeing show a positive relationship. More generally, ES-free descriptions of research findings appear particularly uninformative when people already have sound intuitions about the direction of these effects in the absence of psychology training. Indeed, that laypeople have good knowledge regarding the direction of psychological effects has been claimed by a number of authors (e.g. Houston, 1983; Gordon, Kleiman & Hanie, 1978). The purpose of Study 2 was therefore to gauge how much knowledge laypeople have about the direction of psychological effects. If this knowledge turns out to be rich, the modal textbook strategy of communicating just the direction of

effects would be of little benefit in furthering readers' understanding. However, if this knowledge turns out to be poor, the modal strategy would appear less unsatisfactory; even in the absence of ES information students could still acquire new knowledge about the direction of effects. Indeed, previous research identified a number of systematic misconceptions laypeople hold about psychological phenomena, e.g. that humans use only 10% of their brain (e.g. Bensley, Rainey, Lilienfeld, & Kuehne, 2015). However, our aim here was to test laypeople's intuitions about a *representative* sample of psychological research findings.

Study 2: The Guessability of Central Psychology FindingsMethod

Materials. From our Study 1 sample, we randomly selected 10 passages from each textbook, and turned each into a multiple-choice response questionnaire item. Each item consisted of a passage that introduced the topic and, where necessary, briefly explained relevant concepts. These passages did not provide any clue to the correct answer. Three response options followed, typically capturing the idea that there was no effect, a positive effect, or a negative effect, e.g. "People differ in sensation-seeking, i.e. their drive to search for experiences and feelings that are varied, novel, complex and intense. Which of the following is true? a) High levels of sensation-seeking encourage drinking. b) Low levels of sensation-seeking encourage drinking. c) There is no link between sensation seeking and drinking." (modelled on Berger, 2014, p. 82).

We randomly allocated the 100 items to 20 questionnaires, each containing 5 items with exactly one item from each sub-discipline. After demographic questions, the printed questionnaire illustrated how to address the item, using the size of Earth and Jupiter as an example (Earth bigger; Jupiter bigger; both of equal size). Participants were encouraged to guess whenever they deemed response options equally plausible. Next followed the five items with their response options. The order of items and the order of the three response

options for each item were determined randomly. A debrief sheet was attached, which participants could take home.

Through a copying error, one of the personality items was incomplete and therefore discarded from analysis, which was therefore based on 99 questions.

Participants and procedure. We approached lecturers from various non-psychology undergraduate courses at Northumbria University and asked them for permission to approach students in one of their teaching sessions. Where this could be arranged, we briefly introduced ourselves and our project, before we handed out the questionnaires. Almost all students approached in this way were willing to participate, completed and returned the questionnaires to the researcher, who thanked them for their participation. The use of 20 different questionnaires meant that participants could not compare their answers. Also, the small number of items prevented fatigue (typically, the whole procedure took less than 10 minutes), and we got the impression that participants engaged seriously with the task. Data collection took place from March 2015 until December 2016.

We collected questionnaires from 294 students. Eight participants completed fewer than four questions. Before examining any data, we decided to exclude these participants because they might have been unwilling to properly engage with the task. This left us with 286 participants (seven of whom left one question unanswered): 78% were in the first year of their course, 13% were in second year, and 9% were in third year. Forty-one of our participants studied business studies, six studied creative writing, 14 studied food science, 62 studied law, 44 studied linguistics, and 119 studied sport science. Average age was 20.5 years (SD = 2.8) and 54% of our participants were female.

Results and Discussion

An initial analysis did not suggest systematic performance differences across the six courses participants were enrolled on, $F(5, 288) = 0.83$, $p = .533$; consequently, all further

analyses ignored this factor. The proportion of correct answers for each questionnaire item served as our dependent variable, and we used items as the unit of analysis. Across sub-disciplines, average accuracy was 47% [42%, 51%]. An overview, based on sub-discipline, is provided in Figure 2. As can be seen, participants did not demonstrate accurate intuitions for any of the five sub-disciplines, and only 11 out of 99 items lead to >80% accuracy.

An exploratory ANOVA indicated statistically significant differences between sub-disciplines, $F(4, 94) = 2.49$, $p = .049$. As can be seen from Figure 2, participants' intuitions were about 15% less accurate for findings in biological and social psychology than for either cognitive, developmental, or personality psychology.

Similar studies on laypeople's intuitions about psychological findings (Barnett, 1986; Barnett, Knust, McMillan, Kaufman & Sinisi, 1988; Gordon et al. 1978; Richard, Bond & Stokes-Zoota, 2001) reported substantially better guessing performance, with average percentage correct around 70% (but see Wong, 1995, for an exception). Differences in question format probably played an important role for poor guessing results in our study. The other studies required participants to make a true/false judgement on either the true effect or its opposite (Barnett, 1986; Barnett et al., 1988; Gordon et al., 1978; Richard et al., 2001), or had participants chose between the true effect and its opposite (Wong, 1995); thus, none of these studies included the absence of an effect (H_0 being true) as a response option. However, 90 of our questionnaire items included this option. The questionnaire answer options which presented no effect (e.g. "There is no link between sensation seeking and drinking.") were chosen frequently (377 times; 29% of responses among the 90 relevant questions), but their accuracy (15% correct answers) was substantially below chance performance (33%). Given that participants in the other studies could not express a frequently held but typically erroneous belief in the absence of an effect, we would argue that these studies overestimate the extent to which psychological research findings can be guessed.

Study 1 found that bestselling textbooks typically describe research findings without any indication of ES, i.e. they report only the direction of effects. The results from Study 2 help to better understand the implications of this finding. We argued that the predominant direction-only portrayal of research findings would be particularly uninformative if people, without being trained in psychology, already had reliable knowledge about the direction of these effects. The results from Study 2 clearly suggest that this is not the case. The good news is therefore that even ES-free descriptions of results are typically not without merit, as some have suggested (Houston, 1983; Gordon et al., 1978). Still, the lack of ES information precludes students to develop a fuller understanding of research findings and their implications as we briefly discussed in the context of sleep for adolescents' wellbeing (Gireesh et al., 2018) and as we argue in greater detail below.

Study 2 has obvious limitations. Although the low number of unanswered questions suggests good commitment from participants, they might have performed better if we had incentivised correct responses (Hertwig & Ortmann, 2001).We also relied on textbooks' statements to evaluate the truthfulness of participants' guesses. However, the research that informs these statements might not always be reliable itself (e.g. Hagger et al., 2016; LeBel & Peters, 2011; Open Science Collaboration, 2015). For these reasons, our results might slightly underestimate guessability. Again, we must ask how well our data represent psychology teaching in general. This might be less of a problem here than in Study 1, though. First, the sub-disciplines we selected are very popular on undergraduate programmes in the US (Norcross et al., 2016), and probably beyond. Second, it strikes us as likely that leading textbooks cover, and potentially shape, the content typically addressed in undergraduate teaching.

General Discussion

Analysis of 10 bestselling psychology textbooks from five particularly popular subfields showed that authors typically ignore ES when they portray research findings and only describe the direction of the effect (Study 1). Given that non-experts have scant knowledge regarding the direction of these effects (Study 2), the modal direction-only portrayal of research findings does not appear altogether uninformative. However, we would argue that students would much benefit if their teachers routinely addressed the magnitude of the findings they teach.

This is because ES is often essential for productive reasoning about psychological research results: How strongly various modifiable behaviors are linked to a desirable outcome can aide the search for promising targets for intervention, as we discussed earlier in the context of sleep and adolescents' wellbeing (Gireesh et al., 2018); in order to provide patients with optimal care or to make informed cost-effectiveness decisions, it is necessary to know how effective various treatments are (e.g. Kendall et al., 2016); the variability of ES across contexts might point to important moderators of the investigated effect (e.g. Smith & Glass, 1977); and even the ES of lab experiments, which are sometimes portrayed as meaningless (e.g. Strack, 2017), will typically provide insight about how well the observed effect can be applied in the field, because ESs have been found to correlate substantially between lab and field studies (Mitchell, 2012). Maybe the best argument for the importance of ESs was provided by Jacob Cohen (1994): imagine what we would know about elasticity if experimenters had only ever reported, 'As I pull, it gets longer'! For these and similar reasons, many have called for greater emphasis on ESs in the scholarly communication of research results (e.g. Appelbaum et al., 2018; Cohen, 1994; Cumming, 2014; Ferguson, 2009; Hunter, 1997; Kirk, 1996; Meehl, 1978; Tukey, 1991; Wilkinson & Task Force on Statistical Inference, 1999; Yates, 1951), and with some success (Fritz et al., 2013).

Thus, ESs are a helpful tool to think critically about research. Since this very ability is a central learning goal for psychology students (e.g. American Psychological Association, 2013; Bernstein, 2017), it would seem desirable that students are frequently exposed to ESs. This is because repeated practice and experience with ESs across different contexts should help students to develop relevant expertise (e.g. Ericsson, 2008; Garfield & Ben-Zvi, 2007). To this end, it would be desirable if the teaching of psychological research findings would frequently address the latter's ES (which is, after all, a central aspect of any finding). The results of Study 1 (although focussed on widely-read textbooks and not at teaching in general) do not instil confidence that this happens a lot. Instead, most psychology students might not be particularly exposed to ESs outside their statistics training. Under these circumstances, it is difficult to see how they can develop expertise in the use of this important tool. We therefore encourage tutors to address the magnitude of any research findings they teach. Reassuringly, Goldstein's (2014) textbook demonstrates that it is possible to routinely address ESs and be highly popular at the same time. We discuss the relative merits of three different strategies to communicate ESs.

Three ways to address effect sizes when teaching research results

We discuss the strengths and limitations of three options that require little effort and only basic statistical knowledge. Option one is to report ESs in original units (Baguley, 2009). For categorical data, percentages are useful; our description of Asch's (1956) conformity experiment provides an example. Continuous measures can be compared across groups or conditions via means; e.g. mean weight loss in kg could describe the effectiveness of an intervention programme. The relationship between two continuous measures can be described via regression slopes; e.g., we tell our statistics students how much exam marks, on average, rise for each attended workshop. In our textbook sample, informative ES description

was typically conveyed via original units (81% of *high* passages). ESs in original units are easily computed and easy to understand (provided readers are familiar with the measures), and students do not require statistical training. Obvious limitations are that students will often lack expertise on the relevant measures, and original unit ESs preclude comparisons between studies using different measures.

Our second option, the use of standardised ESs, fixes these problems. (We provide useful code in the Appendix.) Cohen's *d* and Pearson's *r* can be converted into each other. Therefore, one standardised ES suffices to express both, differences between group means and the strength of correlations[2]. We prefer *d* over *r* because only the former is linearly related to the magnitude of effects (for *d*, 0.8 is twice as strong as 0.4, but this is not the case for *r*). A more nuanced discussion of *r* vs. *d* can be found in McGrath and Meyer (2006).

How can we help students to make sense of *d*? We can draw on the probability that a random score from group A is larger than a random score from group B (McGraw & Wong, 1992). Given that people deal more easily with proportions than probabilities (Sedlmeier & Gigerenzer, 2001), it might be even better to point to the proportion of scores in group A that are smaller than the average score in group B (Cohen, 1988). Magnussen (2014) provides a calculator and visualization tool online.[3]

---

[2] Our focus on simple solutions precludes a treatment of factorial designs (see Lakens, 2013, for details). We agree with Ferguson (2009) that results on categorical measures cannot be sensibly transformed into *d* or *r* (Ferguson, 2009), and therefore do not address this point either.

[3] Readers who prefer *r* should consider Rosenthal and Rubin (1982) for lay-friendly illustrations of its meaning.

For an illustration, let us consider secular change in body height. Average height for Dutch conscripts increased by 4.6 cm to 178.7 cm from 1950 to 1970 (van Wieringen, 1986). General familiarity with subject matter and measurement unit make it easy to comprehend this observation. Nonetheless, we might standardise the ES to $d = 0.7$ (SD = 6.7cm in the relevant population). Thus, the same height that was average in 1950 would merely be at the 24th percentile in 1970 (cf. Magnussen, 2014). Importantly, only the standardised ES affords a comparison with other phenomena, e.g. secular change in IQ. For Dutch conscripts, IQ rose by about $d = 0.8$ from 1952 to 1972 (Flynn, 1987). Thus, for the particular population and period, gains in height and IQ were similar.

As drawbacks, standardised ESs require introduction to some elementary statistical concepts. (Maybe this should not be a concern. After all, we cannot imagine physics courses that avoid maths to accommodate students' preferences.) Also, standardised ESs might be somewhat misleading where homogenous samples inflate differences between groups or deflate correlations (Bond, Wiitala, & Richard, 2003). A third option is to describe ESs with easily understood verbal labels (e.g. Berenbaum & Beltz, 2016). Cohen's (1988) convention (*d*s of 0.2, 0.5, and 0.8 are considered *small*, *medium*, and *large*, respectively) has become widely popular. However, if different authors use different conventions (e.g. see Ferguson, 2009, for a far more conservative approach) this will muddy the waters for readers. The vagueness of verbal labels might be seen as an advantage. Compared against the precision of numerical descriptions, verbal labels seem less likely to invoke an unjustified sense of accuracy.

A caveat and concluding remarks

Although we believe that ES is a central aspect of empirical results, which should be considered as much in teaching as in research, a caveat is in order. Like all estimates of population parameters, sample ESs necessarily come with a margin of error, and this is often

considerable. Even close replications of the same study often lead to variability in the ES estimate across samples (Klein et al., 2014). Consequently, uncritical presentation of a point estimate from a single study can be misleading. Estimates from meta-analysis should generally be more trustworthy (Schmidt & Oh, 2016). Where confidence intervals turn out to be wide (or were moderate sample sizes imply this to be the case), it might be appropriate to point out that little is known about the magnitude of the effect.

Although the reporting of (standardised) ESs is on the rise in research publications (Fritz et al., 2013), authors typically fail to address ESs in their discussion (Fritz et al., 2012). This suggests that push for statistical reform has made ES reporting second nature for many authors, whereas their use, for which we have presented a number of examples, has not yet become engrained. A more systematic focus on ES in the teaching of psychology will help foster a generation of future researchers to whom thinking in effect sizes will be natural. Although this is not a panacea (Fiedler, 2017), it will help.

References

American Psychological Association (2010). *Publication manual of the American Psychological Association. 6th ed.* Washington, DC: American Psychological Association.

American Psychological Association (2013). *APA guidelines for the undergraduate psychology major: Version 2.0*. Retrieved from http://www.apa.org/ed/precollege/undergrad/index.aspx

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73(1)*, 3-25.

Aronson, E., & Wilson, T. D. (2012). *Social psychology* (8th ed.). London: Pearson.

Asch, S. E. (1956). Studies of independence and conformity I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*, 1-70.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603–617.

Barnett, M. A. (1986). Commonsense and research findings in personality. *Teaching of Psychology, 13*, 62-64.

Barnett, M. A., Knust, J., McMillan, T., Kaufman, J., & Sinisi. C. (1988). Research findings in developmental psychology: Common sense revisited. *Teaching of Psychology, 15*, 195-197.

Bensley, D. A., Rainey, C., Lilienfeld, S. O., & Kuehne, S. (2015). What do psychology students know about what they know in psychology? *Scholarship of Teaching and Learning in Psychology, 1(4)*, 283-297.Bernstein, D. A. (2017). Bye-bye intro: A proposal for transforming introductory psychology.

Scholarship of Teaching and Learning in Psychology, 3(3), 191-197.

Berenbaum, S. A., & Beltz, A. M. (2016). How early hormones shape gender development. *Current Opinion in Behavioral Sciences, 7*, 53-60.

Berger, K. S. (2014). *The developing person through the life span* (9th ed.). Duffield: Worth Publishers.

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-Analysis of raw mean differences. *Psychological Methods, 8*, 406-418.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Lawrence Erlbaum Associates Publishers: Hillsdale, NJ.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: a general overview. *Academic Emergency Medicine, 15*(11), 988-994.

Feist, J., Feist, G., & Roberts, T.A. (2012). *Theories of personality* (8th ed.). Columbus, OH: McGraw-Hill.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 5*, 532-538.

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science, 12*, 46–61.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

Fritz, C. O., Morris, P.E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2–18.

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting in psychology journals: Are effect sizes really enough? *Theory & Psychology, 23*, 98-122.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372-396.

Gazzaniga, M., Ivry, R. B., & Magnum, G. R. (2013). *Cognitive neuroscience: The biology of the mind* (4th ed.). New York, NY: W. W. Norton & Company.

Gireesh, A., Das, S., & Viner, R. M. (2018). Impact of health behaviours and deprivation on well-being in a national sample of English young people. *BMJ Paediatrics Open, 2(1).*

Goldstein, E. B., (2014). *Cognitive psychology: Connecting mind, research and every day experience* (4th ed.). Independence, KY: Cengage.

Gordon, M. E., Kleiman, L. S., & Hanie, C. A. (1978). Industrial-organizational psychology. Open thy ears o house of Israel. *American Psychologist, 33*, 893-905.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and brain sciences, 24(3)*, 383-403.

Houston, J. P. (1983). Psychology: A closed system of self-evident information? *Psychological Reports, 52*, 203-208.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 3-7.

Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences, 4*, 427-431.

Kalat, J. W. (2012). *Biological psychology* (11th ed.). Independence, KY: Cengage.

Kendall, T., Morriss, R., Mayo-Wilson, E., Meyer, T. D., Jones, S. H., Oud, M., & Baker, M. R. (2016). NICE guidance on psychological treatments for bipolar disorder. *The Lancet Psychiatry, 3*(4), 317-320.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142-152

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4:863.

Larsen, R. J., Buss, D. M., & Wismeijer, A. (2013). *Personality psychology: Domains of knowledge about human nature* (5th ed.). Columbus, OH: McGraw-Hill Higher Education.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*(4), 371-379.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods, 11*, 386-401.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361-365.

Magnussen, K. (2014, February 3). Interpreting Cohen's d effect size. An interactive

visualization. Retrieved from http://rpsychologist.com/d3/cohend/

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*(2), 109-117.

Myers, D. (2012). *Social psychology* (15th ed.). Columbus, OH: McGraw-Hill.

Norcross, J. C., Hailstorks, R., Aiken, L. S., Pfund, R. A., Stamm, K. E., & Christidis, P. (2016). Undergraduate study in psychology: Curriculum and assessment. *American Psychologist, 71*, 89-101.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943-951.Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2001). "That's completely obvious . . .

and important": Lay judgments of social psychological findings. *Personality and Social Psychology Bulletin, 27*, 497-505.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Santrock, J. W. (2013). *Essentials of lifespan development* (3rd ed.). Columbus, OH: McGraw-Hill.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in

psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology,* 4, 32-37.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62*, 97-128.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752-760.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380-400.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487-510.

Sternberg, R. J., & Sternberg, K. (2012). *Cognitive psychology* (6th ed.). Belmont, CA: Wadsworth.

Strack, F. (2017). From data to truth in psychological science. A personal perspective. *Frontiers in Psychology, 8:702*.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

van Wieringen, J. C. (1986). Secular growth changes. In F.T. Falkner & J.M. Tanner (Eds.), *Human Growth: A Comprehensive Treatise* (pp. 307-331.). New York: Plenum Press.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist, 54*, 594-604.

Wong, (1995). Research on teaching: Process-product research findings and the feeling of obviousness. *Journal of Educational Psychology, 87*, 504-511.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.

Table 1. Selection criteria for textbook passages for survey of effect size usage.

Inclusion criteria	Exclusion criteria
- Empirical research on humans. - At least one variable is about behaviour, cognition or similar (e.g. 'serotonin levels rise after eating' but not 'nitric oxide dilates the blood vessels'). - Compares two or more groups or conditions, or uses correlation or regression (thus not 'X% of people in the US are bilingual').	- Insufficiently specific (e.g. 'smoking has detrimental effects on health'). - Textbook states that topic is being dealt with in greater detail elsewhere. - Comparison group only implied but not actually studied (e.g. 'at age X, infants can imitate gestures').

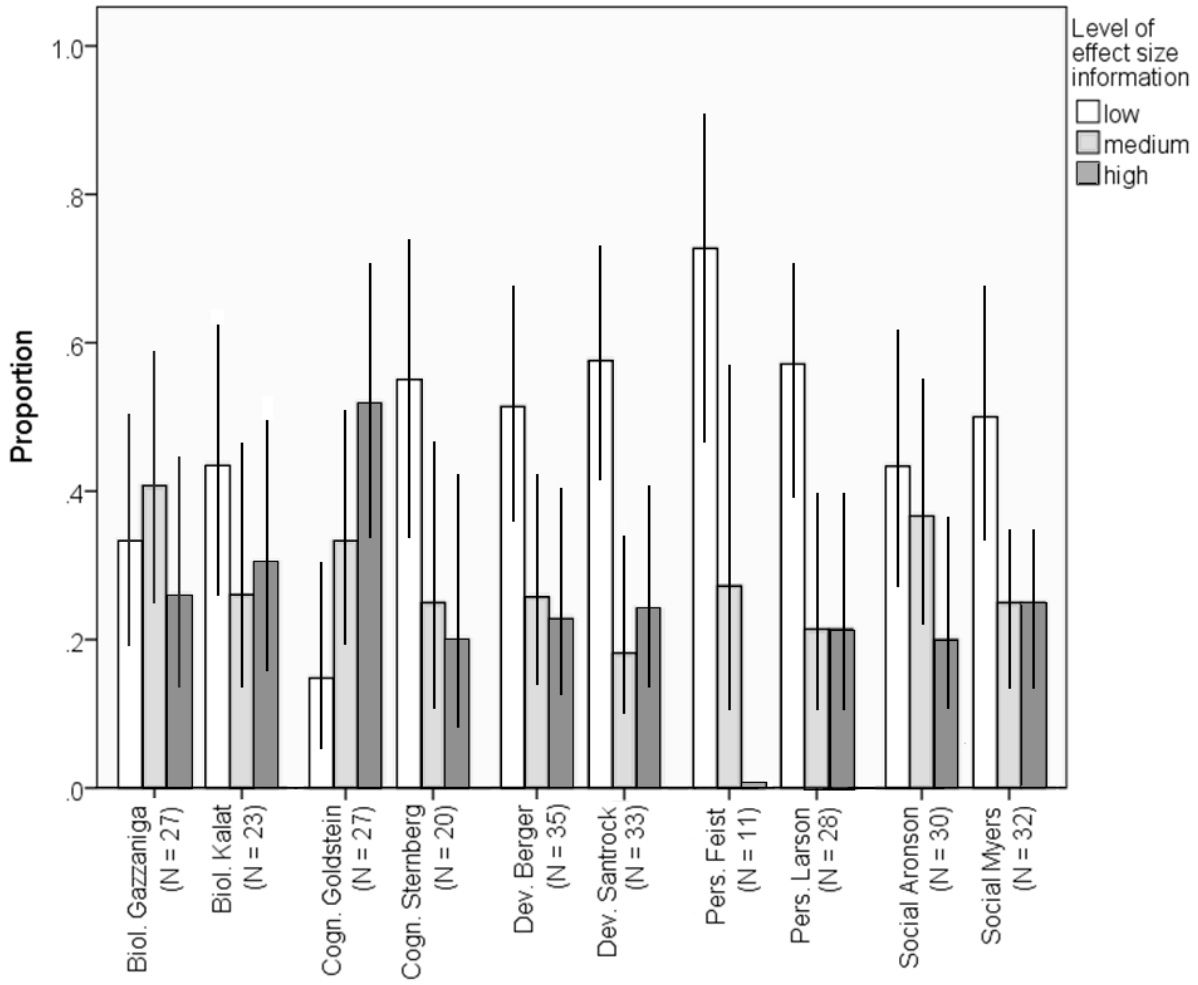

Figure 1. Proportion of low, medium, and high levels of effect size information in passages from 10 textbooks from five psychology sub-fields. *N*s indicate the absolute number of textbook passages; error bars indicate 95% CIs and were calculated with *ESCI effect sizes, proportions* (Cumming, 2013).

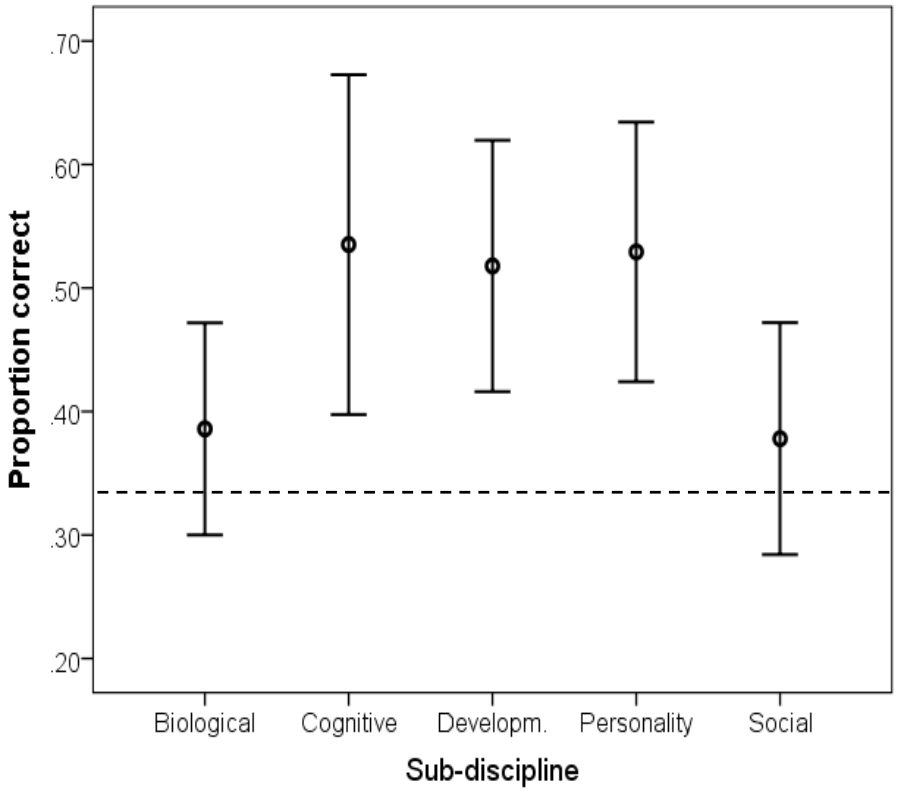

Figure 2. Non-psychology students guessing results for a representative sample of psychology research findings from five sub-disciplines. Mean proportion of correct responses with 95% CIs. Dotted line indicates performance at chance level.