Online Appendix A: additional example of product comparison

The following example uses magnitude-based inference to compare two systems to support the learning of visual programming. The Academy of the Silver Age wants to improve the results of its course (called Hora Est) on introductory computer-programming for older adults. Based on existing work (Vasilopoulos, 2014), the Academy's research team has selected the integrated development environment Koios, to be compared with MicroWorlds Pro, which had previously been used in the course. Koios fully complies with the principles of visualization, support for syntax and semantics, the provision of error messages and a high level of interaction, while MicroWorlds Pro does not (error messages) or partially comply (the other three principles); both environments highly comply with the principles of natural language, abstraction of programming commands and a small set of instructions (Vasilopoulos, 2014). Because Koios fully complies with the seven principles, while MicroWorlds Pro only partially complies, the use of Koios is potentially beneficial.

The research team conducts a field experiment to test the effect of full compliance. A two-group independent measures design is used; each academy member who has signed up for the course is randomly allocated to either MicroWorlds Pro or Koios (*n* = 50 for both) and completes the self-study course, which includes computer-based exercises. Consistent with principles of usability engineering (e.g., Wixon, 2011), the research team sets a measurable product goal in a process of continual improvement: at this stage, after consultation with previous students on the course, the team defines the smallest worthwhile improvement as an increase in programming performance of 10 points on a standardised final practical test, with possible scores from 0 to 100. The mean (± *SD*) practical-test scores were

50.9 ± 9.1 points for MicroWorlds Pro and 63.7 ± 9.0 points for Koios, and the application of magnitude-based inferences enabled the research team to calculate the % chances of benefit (or harm), with reference to a change of 10 points. The following inferences are based on the results of *t* tests, conducted with a common statistics package, which are subsequently used as input into a spreadsheet for magnitude-based inference (see Supplementary Materials A [and *http://sssl-staffweb.tees.ac.uk/U0011128/mbi/*]; main results are presented in Table OA1). In the example of Hora Est, compared to MicroWorlds Pro, performance with Koios was 12.8 points higher (90% confidence interval 9.8 to 15.8), with a 93.9% chance of benefit, a 7.1% chance of negligibility and 0.0% chance of harm; the use of Koios over MicroWorlds Pro is therefore likely to be beneficial and recommended (practical inference) and is likely to be positive and unlikely to be trivial (mechanistic inference).

_____ Insert Table OA1 about here. _____

_____ Interested readers consult Supplementary Material A about here. _____

The University of the Golden Oldies also wants to improve the results of its course (called Tempus Fugit) on introductory computer-programming for older adults. The University's research team employs the same research design as their colleagues at the Academy, but with a different outcome measure. At this stage, after consultation with previous students on the course, the team defines the smallest worthwhile improvement as a decrease in time of 10 days to successful completion of the course. The mean (± *SD*) completion times were 78.3 ± 9.1 days for MicroWorlds Pro and 70.6 ± 11.2 days for Koios, and the application of magnitude-based inferences enabled the research team to calculate the % chances of benefit (or harm), with reference to a change of -10 days. The following inferences are, made with the same approach as in the previous example. In the field experiment of Tempus Fugit, compared to MicroWorlds Pro, completion with Koios was 7.7 days faster (90% confidence interval 3.4 to 11.1), with a 86.5% chance of negligibility and 13.5% chance of benefit; the use of Koios over MicroWorlds Pro is therefore unlikely to be beneficial and not recommended (practical inference) and is likely to be trivial and unlikely to be positive (mechanistic inference).

Note that, if the decision rule 'odds ratio(benefit/harm) > 66' were used to decide on using the effect of Koios, practical inference would be 'use Koios' in both examples – a likely beneficial effect (Hora Est) and a likely trivial effect (Tempus Fugit) – rather than in only the first example (Hora Est). This is because this decision rule does not take into account the absolute size of the probabilities of benefit and harm, whereas the rule that was applied here ('*p*[harm] > 0.25 and *p*[benefit] < 0.005') does. Also note that if NHST were used, the results in both examples would be significant rather than in only one. This is because NHST does not take into account the smallest important effect, whereas magnitude-based inference does.

Reference

Vasilopoulos, I., 2014. The design, development and evaluation of a visual programming tool for novice programmers: psychological and pedagogical effects of introductory programming tools on programming knowledge of Greek students. PhD thesis, Teesside University, Middlesbrough,United Kingdom.

Online Appendix B: exploration of magnitude-based inference through numerical examples

The perceived usability of two product designs (A and B) is tested (Sauro & Lewis, 2012, pp. 69-72). After using one of the designs, each user rates this design with the System Usability Scale (SUS; Sauro, 2011). The results of a *t* test conducted with standard statistical software on the difference of the means between the group using Product Design A and the group using Product Design B are used as input to magnitude-based inference. The chances of positivity/benefit are then calculated as the 100% × p(mean difference > smallest important positive effect) and the chances of positivity/benefit are then calculated as the 100% × p(mean difference < smallest important negative effect), while the chances of triviality/negligibility are calculated as 100% - the other two chances. To help the reader appreciate how the results and conclusions as well as estimated sample size differ between NHST and magnitude inference, Table OA2 presents condensed results for inference. (However, interested readers are also invited to explore the full results presented in Supplementary Material A [and *http://sssl-staffweb.tees.ac.uk/U0011128/mbi/*], tabulated Sheet mean-BS-SUS-1, to observe how magnitude-based inference and NHST compare.) We also encourage interested readers to explore results from their own data sets. They can do this by copying (a datasheet from) the spreadsheet, inserting their own results, experimenting with different researcher-defined smallest important effects and observing how inference changes as a result.

_____ Insert Table OA2 about here. _____

_____ Interested readers consult Supplementary Material A about here. _____

The results are organised in three panels. Within each panel, in each set of results, the following full set of the items (Table OA2 and a superset in Supplementary Material A) are presented in subsequent columns:

- smallest important standardised effect size,
- parameter that is being varied: standardised observed effect size (Table OA2, Panel 1) or sample size (Table OA2, Panels 2 and 3),
- *p*-value (from the results of NHST conducted with a statistical package),
- value of effect size with confidence limits,
- chances for the value of the effect size to be beneficial or substantially positive (as a percentage), negligible or trivial and harmful or substantially negative, and
- inference made.

The qualitative descriptors used in the practical and mechanistic inferences are produced by applying the tabulated matching of probability/chances/odds with adjectives for effects shown in Table 1.[1]

B1. Varying effect size, with constant sample size

In Panel 1, five datasets from the same design are analysed. First, the datasets differ in magnitude of the actual effect of product design, expressed as actual standardised effect size d_{observed} (mean difference/SD_{pooled}), with five values: 1.03, 0.46, -0.11, -0.68, -1.26. Second, three values of the smallest important effect are analysed for each dataset. Degrees of freedom (and thereby sample size) remain constant at 21.

[1] Those who are more interested in some major underlying patterns in the results than in details of the results themselves are invited to skip to Section 5.1.2 in the main text.

In the results for small effect size (*d* = 0.2) as the smallest important effect size, according to NHST, Design A is better than B in the result for $d_{observed}$ = 1.03 and vice versa in the result for $d_{observed}$ = -1.26, but the remaining results are unclear (retain the null hypothesis) and all results remain the same irrespectively of smallest important effect size (*d* = 0.2, *d* = 0.6, *d* = 1.2). However, the changes in practical and mechanistic inference reflect a decrease in chances for benefit or substantially positivity and an increase for harm or substantial negativity when we move from $d_{observed}$ = 1.03 to 0.46 to -0.11 to -0.68 to -1.26, with the chances of a negligible or trivial effect most pronounced for $d_{observed}$ = 0.46 and -0.11. In particular, according to magnitude-based practical inference, only the result for $d_{observed}$ = 0.46 is unclear; otherwise Design A should be used ("very likely beneficial, most unlikely harmful" [$d_{observed}$ = 1.03]) or Design A should be not used ("possibly harmful, unlikely beneficial" [$d_{observed}$ = -0.11], "likely harmful, very unlikely beneficial" [$d_{observed}$ = -0.68] and "very likely harmful, most unlikely beneficial" [$d_{observed}$ = -1.26]). Moreover, according to magnitude-based mechanistic inference, only the results for $d_{observed}$ = 0.46 and $d_{observed}$ = -0.11 are unclear, and the difference in favour of Design A is "very likely positive" ($d_{observed}$ = 1.03), "likely negative" ($d_{observed}$ = -0.68) or "very likely negative" ($d_{observed}$ = -1.26).

For a moderate effect size (*d* = 0.6), all results of statistical inference remain the same as before. However, again, the changes in practical and mechanistic inference reflect a decrease in chances for benefit or substantially positivity and an increase for harm or substantial negativity when we move from $d_{observed}$ = 1.03 to 0.46 to -0.11 to -0.68 to -1.26, with the chances of a negligible or trivial effect most pronounced for $d_{observed}$ = 0.46, -0.11 and -0.68, and the following results reflect this. Moreover, compared with the results for a small effect size (*d* = 0.2), we can see in the probabilities for the three ranges of effect (beneficial or substantially positive, negligible or trivial, harmful or substantially negative) that those in the middle range increase and those in the other ranges decrease. As a consequence, the results of magnitude-based practical inference have changed (from those for a small effect size, *d* = 0.2) in terms of their qualitative description: the result for $d_{observed}$ = 0.46 is unclear; otherwise Design A should be used (“likely beneficial, most unlikely harmful”, $d_{observed}$ = 1.03) or Design A should be not used (“unlikely harmful, unlikely beneficial” [$d_{observed}$ = -0.11], “possibly harmful, most unlikely beneficial” [$d_{observed}$ = -0.68] and “likely harmful, most unlikely beneficial” [$_{dobserved}$ = -1.26]). Moreover, the results for mechanistic inference have changed in terms of clarity of results and their qualitative description: the result for $d_{observed}$ = -0.11 is unclear; otherwise, the difference in favour of Design A is “likely positive” ($_{dobserved}$ = 1.03), “possibly trivial” ($d_{observed}$ = 0.46), “possibly negative” ($d_{observed}$ = -0.68) or “likely negative” ($d_{observed}$ = -1.26).

For a large effect size (*d* = 1.2), all results of statistical inference remain the same as before. However, again this time, the changes in practical and mechanistic inference reflect a decrease in chances for benefit or substantially positivity and an increase for harm or substantial negativity when we move from $d_{observed}$ = 1.03 to 0.46 to -0.11 to -0.68 to -1.26; moreover, the chances of a negligible or trivial effect are substantial for all observed effects, most notably for $d_{observed}$ = 0.46, -0.11 and -0.68, and the following results reflect this. Furthermore, compared with the results for small and moderate effect sizes (*d* = 0.2 and *d* = 0.6, respectively) we can see in the probabilities for the three ranges of effect (beneficial or substantially positive, negligible or trivial, harmful or substantially negative) than those in the middle range further increase and those in the other ranges further decrease. As a consequence, for a large effect size (*d* = 1.2), the results of magnitude-based practical inference have changed again (from those for a moderate effect size, *d* = 0.6) in terms of clarity of results and their qualitative description: all results are clear, and Design A should be used ("possibly beneficial, most unlikely harmful" [$d_{observed}$ = 1.03] and "very unlikely beneficial, most unlikely harmful" [$d_{observed}$ = 0.46]) or Design A should be not used ("very unlikely harmful, most unlikely beneficial" [$d_{observed}$ = -0.11], "unlikely harmful, most unlikely beneficial" [$d_{observed}$ = -0.68] and "possibly harmful, most unlikely beneficial" [$d_{observed}$ = -1.26]). Moreover, the results for mechanistic inference have changed in terms of clarity of results and their qualitative description: all results are clear, and the difference in favour of Design A is "possibly trivial" ($_{dobserved}$ = 1.03), "very likely trivial" ($d_{observed}$ = 0.46), "very likely trivial" ($d_{observed}$ = -0.11), "likely trivial" ($d_{observed}$ = -0.68) or "possibly negative" ($d_{observed}$ = -1.26).

B2. Varying sample size, with constant *p*-value

In Panel 2, again five datasets from the same design are analysed. First, a combination of varying sample size with fixed *p*-value (0.12) is achieved through decreasing effect size: *N* = 23 with *d* = -0.68, *N* = 46 with *d* = -0.48, *N* = 69 with *d* = -0.39, *N* = 92 with *d* = -0.34 and *N* = 115 with *d* = -0.30. Second, different values of the smallest important effect are analysed for each dataset. Degrees of freedom (and thereby sample size) varies as a function of sample size.

In the results for small effect size (*d* = 0.2) as the smallest important effect size, according to NHST, all results are unclear (retain the null hypothesis) and remain the same irrespectively of smallest important effect size (*d* = 0.2, *d* = 0.6, *d* = 1.2). However, with increasing sample size (and decreasing effect size, together producing a constant *p*-value) the chances of benefit/substantial positivity and harm/substantial negativity decrease and the chances of a negligible or trivial effect increase and the following results of practical and mechanistic inference reflect this. According to magnitude-based practical inference, all results are clear, and Design A should not be used ("likely harmful, very unlikely beneficial" [*N* = 23, *N* = 46, *N* = 69]; "possibly harmful, very unlikely beneficial" [*N* = 92, *N* = 115]). Moreover, according to magnitude-based mechanistic inference, all results are clear as well, and the difference in favour of Design A is "likely negative" (*N* = 23, *N* = 46, *N* = 69) or "possibly negative" (*N* = 92, *N* = 115).

For a moderate effect size (*d* = 0.6), all results of statistical inference remain unclear (retain the null hypothesis) and the same as before. However, again with increasing sample size (and decreasing effect size, together producing a constant *p*-value) the chances of benefit/substantial positivity and harm/substantial negativity decrease and the chances of a negligible or trivial effect increase even more this time, and the following results of practical and mechanistic inference reflect this. For a moderate effect size (*d* = 0.6), the results of magnitude-based practical inference have only changed (from those for a small effect size, *d* = 0.2) in terms of their qualitative description: Design A should not be used ("possibly harmful, most unlikely beneficial" [*N* = 23, *N* = 46]; "unlikely harmful, most unlikely beneficial" [*N* = 69, *N* = 92, *N* = 115]). Moreover, the results for mechanistic inference have changed in terms of their qualitative description: the difference in favour of Design A is "possibly negative" (*N* = 23), "possibly trivial" (*N* = 46), "likely trivial" (*N* = 69, *N* = 92, *N* = 115).

For a large effect size (*d* = 1.2), all results of statistical inference remain unclear (retain the null hypothesis) and the same as before. However, this time again with increasing sample size (and decreasing effect size, together producing a constant *p*-value) the (very small) chances of benefit/substantial positivity and harm/substantial negativity decrease even further and the very large chances of a negligible or trivial effect increase to even 100%, and the following results of practical and mechanistic inference reflect this. The results of magnitude-based practical inference have only changed (from those for a moderate effect size, *d* = 0.6) in terms of their qualitative description: Design A should not be used ("unlikely harmful, most unlikely beneficial" [*N* = 23]; "very unlikely harmful, most unlikely beneficial" [N = 46]; "most unlikely harmful, most unlikely beneficial" [*N* = 69, *N* = 92, *N* = 115]). Moreover, the results for mechanistic inference have changed in terms of their qualitative description: the difference in favour of Design A is "likely trivial" (*N* = 23), "very likely trivial" (*N* = 46), "most likely trivial" (*N* = 69, *N* = 92, *N* = 115).

B3. Varying sample size, with constant effect size

In Panel 3, again five datasets from the same design are analysed. First, a combination of varying sample size with fixed raw effect size (mean difference = 2.01 [and $d_{observed}$ approximately constant at 0.46 to 0.48]) is achieved through decreasing *p*-value: *N* = 23 with *p* = 0.28, *N* = 46 with *p* = 0.12, *N* = 92 with *p* = 0.03, *N* = 184 with *p* = 0.001 and *N* = 368 with *p* = 0.000006. Second, different values of the smallest important effect are analysed for each dataset. Degrees of freedom (and thereby sample size) varies as a function of sample size.

In the results for small effect size (*d* = 0.2) as the smallest important effect size, according to NHST, Design A is better than B in the result for *N* = 92, *N* = 184 and *N* = 368, with decreasing *p*-value, but the remaining results (for *N* = 23 and *N* = 46) are unclear (retain the null hypothesis) and all results remain the same irrespectively of smallest important effect size (*d* = 0.2, *d* = 0.6, *d* = 1.2). However, as the precision of inference increases with sample size, for *d* = 0.2 the chances of benefit or substantial positivity increase (as the actual observed effect at approximately 0.47 is greater than 0.2) and the chances of a negligible or trivial effect and of harm or substantial negativity decrease, and this is reflected in the following results. According to magnitude-based practical inference, the results are clear except when *N* = 23, and Design A should be used ("likely beneficial, very unlikely harmful" [*N* = 46]; "likely beneficial, most unlikely harmful" [*N* = 92]; "very likely beneficial, most unlikely harmful" [*N* = 184]; "mostly likely beneficial, most unlikely harmful" [*N* = 368]). Moreover, according to magnitude-based mechanistic inference, all results are clear as well except when *N* = 23, and the difference in favour of Design A is "likely positive" (*N* = 46, *N* = 92), "very likely positive" (*N* = 184) or "most likely positive" (*N* = 368).

For a moderate effect size (*d* = 0.6), all results of statistical inference remain the same as before. However, as the precision of inference increases with sample size, compared with the results for a small effect size (*d* = 0.2), for *d* = 0.6 the chances of a negligible or trivial effect increase (as the actual observed effect at approximately 0.47 is smaller than 0.6) and the chances of benefit or substantial positivity (and of harm or substantial negativity) decrease, and this is reflected in the following results. The results of magnitude-based practical inference have only changed (from those for a small effect size, *d* = 0.2) in terms of their qualitative description: except for *N* = 23, when the result is unclear, Design A should be used ("possibly beneficial, most unlikely harmful" [*N* = 46]; "unlikely beneficial, most unlikely harmful" [*N* = 92, *N* = 184, *N* = 368]). Moreover, the results for mechanistic inference have changed in terms of their qualitative description: the difference in favour of Design A is "possibly trivial" (*N* = 23, *N* = 46) or "likely trivial" (*N* = 92, *N* = 184, *N* = 368).

For a large effect size (*d* = 1.2), all results of statistical inference remain the same as before. However, because the actual observed effect at approximately 0.47 is considerably smaller than 1.2, the chances of a negligible or trivial effect are extremely high (95% to 100%), and this is reflected in the following results. The results of magnitude-based practical inference have changed (from those for a moderate effect size, *d* = 0.6) in terms of clarity and their qualitative description: all results are clear, and Design A should be used ("very unlikely beneficial, most unlikely harmful" [*N* = 23, *N* = 46]; "most unlikely beneficial, most unlikely harmful" [*N* = 92, *N* = 184, *N* = 368]). Moreover, the results for mechanistic inference have changed in terms of clarity and their qualitative description: all results are clear, and the difference in favour of Design A is "very likely trivial" (*N* = 23, *N* = 46) or "most likely trivial" (*N* = 92, *N* = 184, *N* = 368).

Reference

Sauro, J., Lewis, J.R., 2012. Quantifying the User Experience. Morgan Kaufmann, Amsterdam.

Table OA1

Inference for between-subjects design

a. Outcome: practical-test score (Hora Est)

10-point difference on practical test as the smallest important effect size

$d_{observed}$	*p*-value	Outcome	Chances that the true value of the effect statistic is …				Inference	Odds ratio (B/H)
		value, 90% CI	beneficial or substantially +ive	negligible or trivial	harmful or substantially -ive			
1.42	2.1E-10	12.8, 9.8 to 15.8	93.9 %	6.1 %	0.0 %	P	likely beneficial, most unlikely harmful; use	1.01×10^{23}
			likely	unlikely	most unlikely	M	likely +ive	
						S	significant	

b. Outcome: time to complete course successfully (Tempus Fugit)

10-point difference on time to complete (days) as the smallest important effect size

$d_{observed}$	*p*-value	Outcome	Chances that the true value of the effect statistic is …				Inference	
		value, 90% CI	beneficial or substantially +ive	negligible or trivial	harmful or substantially -ive			
0.76	0.00027	7.7, 4.3 to 11.1	13.5 %	86.5 %	0.0 %	P	likely negligible; don’t use	3.35×10^{12}
			likely	unlikely	most unlikely	M	likely trivial	
						S	significant	

Note. d: standardised mean difference. d = 0.2 as the smallest important effect size. P: practical inference. M: mechanistic inference. S: statistical inference (NHST). B: benefit. H: harm.

Table OA2

Inference for between-subjects design (perceived usability)

Panel 1: varying effect size, with constant sample size (N = 22 = 11 + 11 = $n_1 + n_2$)
Data adapted and expanded from Sauro and Lewis, Example 2, pp. 69-72

d = 0.2 as the smallest important effect size

$d_{observed}$	*p*-value	Outcome	Chances that the true value of the effect statistic is …							Inference
			beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
		value, 90% CI								
1.03	0.02	4.5, 1.4 to 7.6	97.02	%	2.60	%	0.38	%	P	very likely beneficial, most unlikely harmful; use
			very likely		very unlikely		most unlikely		M	very likely +ive
									S	significant
0.46	0.28	2, -1.1 to 5.1	72.94	%	20.60	%	6.46	%	P	Unclear; don't use; get more data
			possibly		unlikely		unlikely		M	unclear; get more data
									S	non-significant
-0.11	0.79	-0.49, -3.6 to 2.6	23.11	%	35.08	%	41.81	%	P	possibly harmful, unlikely beneficial; don't use
			unlikely		possibly		possibly		M	unclear; get more data
									S	non-significant
-0.68	0.12	-3, -6.1 to 0.15	2.31	%	10.63	%	87.06	%	P	likely harmful, very unlikely beneficial; don't use
			very unlikely		unlikely		likely		M	likely -ive
									S	non-significant
-1.26	0.01	-5.5, -8.6 to -2.4	0.11	%	0.86	%	99.03	%	P	very likely harmful, most unlikely beneficial; don't use
			most unlikely		very unlikely		very likely		M	very likely -ive
									S	significant

Note. d: standardised mean difference. P: practical inference. M: mechanistic inference. S: statistical inference (NHST).

Table OA2, Panel 1 (continued)

	d = 0.6 as the smallest important effect size									
$d_{observed}$	*p*-value	Outcome	Chances that the true value of the effect statistic is … Inference							
		value, 90% CI	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
1.03	0.02	4.5, 1.4 to 7.6	84.33	%	15.63	%	0.04	%	P	likely beneficial, most unlikely harmful; use
			likely		unlikely		most unlikely		M	likely +ive
									S	significant
0.46	0.28	2, -1.1 to 5.1	36.96	%	62.08	%	0.96	%	P	Unclear; don't use; get more data
			possibly		possibly		very unlikely		M	possibly trivial
									S	non-significant
-0.11	0.79	-0.49, -3.6 to 2.6	5.13	%	82.07	%	12.80	%	P	unlikely harmful, unlikely beneficial; don't use
			unlikely		likely		unlikely		M	unclear; get more data
									S	non-significant
-0.68	0.12	-3, -6.1 to 0.15	0.29	%	41.79	%	57.93	%	P	possibly harmful, most unlikely beneficial; don't use
			most unlikely		possibly		possibly		M	possibly -ive
									S	non-significant
-1.26	0.01	-5.5, -8.6 to -2.4	0.01	%	6.53	%	93.46	%	P	likely harmful, most unlikely beneficial; don't use
			most unlikely		unlikely		likely		M	likely -ive
									S	significant

Table OA2, Panel 1 (continued)

	d = 1.2 as the smallest important effect size									
$d_{observed}$	*p*-value	Outcome	Chances that the true value of the effect statistic is …							Inference
		value, 90% CI	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
1.03	0.02	4.5, 1.4 to 7.6	34.49	%	65.51	%	0.00	%	P	possibly beneficial, most unlikely harmful; use
			possibly		possibly		most unlikely		M	possibly trivial
									S	significant
0.46	0.28	2, -1.1 to 5.1	4.52	%	95.44	%	0.03	%	P	unclear; don’t use; get more data
			very unlikely		very likely		most unlikely		M	very likely trivial
									S	non-significant
-0.11	0.79	-0.49, -3.6 to 2.6	0.24	%	98.93	%	0.83	%	P	very unlikely harmful, most unlikely beneficial; don't use
			most unlikely		very likely		very unlikely		M	very likely trivial
									S	non-significant
-0.68	0.12	-3, -6.1 to 0.15	0.01	%	88.47	%	11.53	%	P	unlikely harmful, most unlikely beneficial; don't use
			most unlikely		likely		unlikely		M	likely trivial
									S	non-significant
-1.26	0.01	-5.5, -8.6 to -2.4	0.00	%	44.69	%	55.31	%	P	possibly harmful, most unlikely beneficial; don't use
			most unlikely		possibly		possibly		M	possibly -ive
									S	significant

Table OA2 (continued)

Panel 2: varying sample size, with approximately constant *p*-value
Data inspired by Sauro and Lewis (2012), Example 2, pp. 69-72

d = 0.2 as the smallest important effect size

N	*p*-value	Outcome	Chances that the true value of the effect statistic is …							Inference
		value, 90% CI	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
23	0.12	-3, -6.1 to 0.15	2.31	%	10.63	%	87.06	%	P	likely harmful, very unlikely beneficial; don't use
			very unlikely		unlikely		likely		M	likely -ive
									S	non-significant
46	0.12	-2, -4.1 to 0.13	1.50	%	16.72	%	81.78	%	P	likely harmful, very unlikely beneficial; don't use
			very unlikely		unlikely		likely		M	likely -ive
									S	non-significant
69	0.12	-1.6, -3.2 to 0.091	1.00	%	21.20	%	77.80	%	P	likely harmful, very unlikely beneficial; don't use
			very unlikely		unlikely		likely		M	likely -ive
									S	non-significant
92	0.12	-1.3, -2.8 to 0.07	0.70	%	25.10	%	74.20	%	P	possibly harmful, very unlikely beneficial; don't use
			very unlikely		possibly		possibly		M	possibly -ive
									S	non-significant
115	0.12	-1.2, -2.4 to 0.078	0.54	%	29.65	%	69.81	%	P	possibly harmful, very unlikely beneficial; don't use
			very unlikely		possibly		possibly		M	possibly -ive
									S	non-significant

Table OA2, Panel 2 (continued)

d = 0.6 as the smallest important effect size

N	*p*-value	Outcome	Chances that the true value of the effect statistic is …							Inference
		value, 90% CI	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
23	0.12	-3, -6.1 to 0.15	0.29	%	41.79	%	57.93	%	P	possibly harmful, most unlikely beneficial; don't use
			most unlikely		possibly		possibly		M	possibly -ive
									S	non-significant
46	0.12	-2, -4.1 to 0.13	0.04	%	65.74	%	34.21	%	P	possibly harmful, most unlikely beneficial; don't use
			most unlikely		possibly		possibly		M	possibly trivial
									S	non-significant
69	0.12	-1.6, -3.2 to 0.091	0.01	%	79.88	%	20.11	%	P	unlikely harmful, most unlikely beneficial; don't use
			most unlikely		likely		unlikely		M	likely trivial
									S	non-significant
92	0.12	-1.3, -2.8 to 0.07	0.00	%	88.40	%	11.60	%	P	unlikely harmful, most unlikely beneficial; don't use
			most unlikely		likely		unlikely		M	likely trivial
									S	non-significant
115	0.12	-1.2, -2.4 to 0.078	0.00	%	93.82	%	6.18	%	P	unlikely harmful, most unlikely beneficial; don't use
			most unlikely		likely		unlikely		M	likely trivial
									S	non-significant

Table OA2, Panel 2 (continued)

d = 1.2 as the smallest important effect size

N	*p*-value	Outcome	Chances that the true value of the effect statistic is …							Inference
		value, 90% CI	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
23	0.12	-3, -6.1 to 0.15	0.01	%	88.47	%	11.53	%	P	unlikely harmful, most unlikely beneficial; don't use
									M	likely trivial
									S	non-significant
46	0.12	-2, -4.1 to 0.13	0.00	%	98.96	%	1.04	%	P	very unlikely harmful, most unlikely beneficial; don't use
			most unlikely		very likely		very unlikely		M	very likely trivial
									S	non-significant
69	0.12	-1.6, -3.2 to 0.091	0.00	%	99.91	%	0.09	%	P	most unlikely harmful, most unlikely beneficial; don't use
			most unlikely		most likely		most unlikely		M	most likely trivial
									S	non-significant
92	0.12	-1.3, -2.8 to 0.07	0.00	%	99.99	%	0.01	%	P	most unlikely harmful, most unlikely beneficial; don't use
			most unlikely		most likely		most unlikely		M	most likely trivial
									S	non-significant
115	0.12	-1.2, -2.4 to 0.078	0.00	%	100.00	%	0.00	%	P	most unlikely harmful, most unlikely beneficial; don't use
			most unlikely		most likely		most unlikely		M	most likely trivial
									S	non-significant

Table OA2 (continued)

Panel 3: varying sample size, with approximately constant effect size ($d \approx 0.47$)
Statistics as input inspired by Sauro and Lewis (2012), Example 2, pp. 69-72

d = 0.2 as the smallest important effect size

N	*p*-value	Outcome expressed as either…	Chances that the true value of the effect statistic is …							Inference
		value, with 90% confidence interval	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
23	0.28	2, -1.1 to 5.1	72.94	%	20.60	%	6.46	%	P	Unclear; don't use; get more data
			possibly		unlikely		unlikely		M	unclear; get more data
									S	non-significant
46	0.12	2, -0.11 to 4.1	81.74	%	16.85	%	1.41	%	P	likely beneficial, very unlikely harmful; use
			likely		unlikely		very unlikely		M	likely +ive
									S	non-significant
92	0.03	2, 0.54 to 3.5	90.48	%	9.44	%	0.08	%	P	likely beneficial, most unlikely harmful; use
			likely		unlikely		most unlikely		M	likely +ive
									S	significant
184	0.001	2, 0.98 to 3	96.94	%	3.06	%	0.00	%	P	very likely beneficial, most unlikely harmful; use
			very likely		very unlikely		most unlikely		M	very likely +ive
									S	significant
368	0.000006	2, 1.3 to 2.7	99.61	%	0.39	%	0.00	%	P	most likely beneficial, most unlikely harmful; use
			most likely		most unlikely		most unlikely		M	most likely +ive
									S	significant

Table OA2, Panel 3 (continued)

	d = 0.6 as the smallest important effect size									
N	*p*-value	Outcome expressed as either…	Chances that the true value of the effect statistic is …							Inference
		value, with 90% confidence interval	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
23	0.28	2, -1.1 to 5.1	36.96	%	62.08	%	0.96	%	P	Unclear; don't use; get more data
			possibly		possibly		very unlikely		M	possibly trivial
									S	non-significant
46	0.12	2, -0.11 to 4.1	31.39	%	68.57	%	0.03	%	P	possibly beneficial, most unlikely harmful; use
			possibly		possibly		most unlikely		M	possibly trivial
									S	non-significant
92	0.03	2, 0.54 to 3.5	24.34	%	75.66	%	0.00	%	P	likely trivial; don’t use
			unlikely		likely		most unlikely		M	likely trivial
									S	significant
184	0.001	2, 0.98 to 3	16.11	%	83.89	%	0.00	%	P	likely trivial; don’t use
			unlikely		likely		most unlikely		M	likely trivial
									S	significant
368	0.000006	2, 1.3 to 2.7	8.00	%	92.00	%	0.00	%	P	likely trivial; don’t use
			unlikely		likely		most unlikely		M	likely trivial
									S	significant

Table OA2, Panel 3 (continued)

	d = 1.2 as the smallest important effect size									
N	*p*-value	Outcome expressed as either…	Chances that the true value of the effect statistic is …							Inference
		value, with 90% confidence interval	beneficial or substantially +ive		negligible or trivial		harmful or substantially -ive			
23	0.28	2, -1.1 to 5.1	4.52	%	95.44	%	0.03	%	P	likely trivial; don’t use
			very unlikely		very likely		most unlikely		M	very likely trivial
									S	non-significant
46	0.12	2, -0.11 to 4.1	0.87	%	99.13	%	0.00	%	P	likely trivial; don’t use
			very unlikely		very likely		most unlikely		M	very likely trivial
									S	non-significant
92	0.03	2, 0.54 to 3.5	0.04	%	99.96	%	0.00	%	P	likely trivial; don’t use
			most unlikely		most likely		most unlikely		M	most likely trivial
									S	significant
184	0.001	2, 0.98 to 3	0.00	%	100.00	%	0.00	%	P	likely trivial; don’t use
			most unlikely		most likely		most unlikely		M	most likely trivial
									S	significant
368	0.000006	2, 1.3 to 2.7	0.00	%	100.00	%	0.00	%	P	likely trivial; don’t use
			most unlikely		most likely		most unlikely		M	most likely trivial
									S	significant