Denouncing the use of field-specific effect size distributions to inform magnitude

Emily Panzarella [Corresp., 1] , **Nataly Beribisky** [1] , **Robert A Cribbie** [1]

[1] Department of Psychology, York University, Toronto, Ontario, Canada

Corresponding Author: Emily Panzarella
Email address: emilyp99@my.yorku.ca

An effect size (ES) provides valuable information regarding the magnitude of effects, with the interpretation of magnitude being the most important. Interpreting ES magnitude requires combining information from the numerical ES value and the context of the research. However, many researchers adopt popular benchmarks such as those proposed by Cohen. More recently, researchers have proposed interpreting ES magnitude relative to the distribution of observed ESs in a specific field, creating unique benchmarks for declaring effects small, medium or large. However, there is no valid rationale whatsoever for this approach. This study was carried out in two parts: 1) We identified articles that proposed the use of field-specific ES distributions to interpret magnitude (primary articles); and 2) We identified articles that cited the primary articles and classified them by year and publication type. The first type consisted of methodological papers. The second type included articles that interpreted ES magnitude using the approach proposed in the primary articles. There has been a steady increase in the number of methodological and substantial articles discussing or adopting the approach of interpreting ES magnitude by considering the distribution of observed ES in that field, even though the approach is devoid of a theoretical framework. It is hoped that this research will restrict the practice of interpreting ES magnitude relative to the distribution of ES values in a field and instead encourage researchers to interpret such by considering the specific context of the study.

1 Denouncing the Use of Field-Specific Effect Size
2 Distributions to Inform Magnitude
3

5   Emily Panzarella[1], Nataly Beribisky[1], Robert A. Cribbie[1]

7   [1] Department of Psychology, York University, Toronto, Ontario, Canada

9   Corresponding Author:
10 Emily Panzarella[1]
11 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada
12 Email address: emilyp99@my.yorku.ca

14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

Abstract

An effect size (ES) provides valuable information regarding the magnitude of effects, with the interpretation of magnitude being the most important. Interpreting ES magnitude requires combining information from the numerical ES value and the context of the research. However, many researchers adopt popular benchmarks such as those proposed by Cohen. More recently, researchers have proposed interpreting ES magnitude relative to the distribution of observed ES values in a specific field, creating unique benchmarks for declaring effects small, medium or large. However, there is no valid rationale whatsoever for this approach. This study was carried out in two parts: 1) We identified articles that proposed the use of field-specific ES distributions to interpret magnitude (*primary articles*); and 2) We identified articles that cited the primary articles and classified them by year and publication type. The first type consisted of methodological papers. The second type included articles that interpreted ES magnitude using the approach proposed in the primary articles. There has been a steady increase in the number of methodological and substantial articles discussing or adopting the approach of interpreting ES magnitude by considering the distribution of ESs in that field, even though the approach is devoid of a theoretical framework. It is hoped that this research will restrict the practice of interpreting ES magnitude relative to the distribution of ES values in a field and instead encourage researchers to interpret such by considering the specific context of the study.

Introduction

An effect size (ES) is defined as a quantitative representation of an event which assesses a research question, often reported alongside a *p*-value (Kelley & Preacher, 2012). Whereas *p*-values only assess whether or not an effect is statistically significant, an ES provides valuable information regarding the strength and direction of relationships among variables. This information is conveyed through an interpretation of an ES value's magnitude. The recommended approach for interpreting the magnitude of effects is to consider the context of the research; however, the method of choice for interpreting the magnitude of an ES is often published benchmarks, such as those proposed by Cohen (1988). According to Google (2021), Cohen's book, which proposes (small, medium, and large) cut-offs for most popular ES measures, has been cited almost 200,000 times!

Recently, researchers have proposed interpreting the magnitude of an ES by considering where it falls relative to the distribution of an observed ES within the specific field in which the work was conducted (e.g., Hemphill, 2003). This approach involves establishing unique benchmark values for interpreting ES magnitude based on the observed ES from sets of studies carried out in a specific field. However, this practice completely ignores the context of the study, which is a critical component when evaluating the meaningfulness of effects. Although some might argue that using the distribution of ES values in a field is *taking context into account*, we completely disagree with this contention. The context of the study considers factors or circumstances directly relevant to the research question(s), such as prior research, setting,

82 participants, methods, and analyses. In contrast, the benchmark approach based on observed ESs
83 necessarily considers only a general field or subfield.
84   In this study, we systematically review the popularity of using the observed ES
85 distribution to interpret ES magnitude in psychological research. Further, we discuss the
86 limitations of the approach and provide recommendations for interpreting ES magnitude in a
87 meaningful and appropriate manner.

88 **Reporting Effect Size**

89   An ES can offer information about the calculated value of effects, the direction of effects,
90 the magnitude of effects, and relevant interpretations of effects (Nakagawa & Cuthill, 2007).
91 Organizations such as *The American Psychological Association* (APA), *The American*
92 *Educational Research Association* (AERA), *The Society for Industrial and Organizational*
93 *Psychology* (SIOP), and *The National Centre for Education Statistics* (NCES) have released
94 official statements regarding the importance of incorporating ES values into research (Kelley &
95 Preacher, 2012). In turn, almost all journal editors have required that an ES value, and its
96 corresponding confidence interval, be reported alongside, or instead of, *p*-values and null
97 hypothesis significance test results. A good ES value should be scaled accordingly given the
98 question of interest, be accompanied by a confidence interval, and be asymptotically unbiased,
99 consistent, and efficient (Kelley & Preacher, 2012).
100   An ES measure can be expressed in either unstandardized or standardized format, with
101 the former representing magnitude in the units of the study variables and the latter representing
102 magnitude in generic units. Examples of unstandardized ES measures includes mean differences
103 and regression coefficients, whereas examples of standardized ES measures include standardized
104 mean differences, correlation coefficients, standardized regression coefficients, and odds/risk
105 ratios (Kelley & Preacher, 2012; Rosenthal et al., 1994). Standardized mean differences are often
106 categorized as part of the *d* family of ES measures (e.g., Cohen's *d*, Hedges' *g*: a bias-corrected
107 version of Cohen's *d*, Glass' Δ: a version of Cohen's *d* which uses a single group's standard
108 deviation), whereas correlation coefficients are often categorized as part of the *r* family of ES
109 measures (e.g., Pearson's *r*, partial/semi-partial *r*, multiple *R*, biserial/point-biserial *r*). Although
110 *d* was designed as a measure of standardized mean difference and *r* was designed as a measure of
111 correlation, these measures can be converted between one another in order to facilitate
112 comparisons across settings and designs (e.g., *d* to *r*, *r* to *d*, odds ratio to *d*, etc.).
113   As constructs become more abstract, if they are not measured reliably, standardized ES
114 measures are at risk of increased measurement error. Increased measurement error results in
115 lower standardized ES values (Burchinal, 2008). Although there are adjustments that may be
116 used for reliability that may produce potentially better estimates of ES (Baguley, 2009), there is
117 no substitute for better measurement and ongoing validation of the construct itself (Flake et al.,
118 2017).
119   ES values results from the application of a relevant ES measure to given data (Kelley &
120 Preacher, 2012). For example, the correlation between two measures can be expressed with an
121 ES value of r = 0.78. Alternatively, an unstandardized mean difference between two groups on

122   the number of cookies consumed can be expressed with an ES value of, for example, 2 cookies.
123   However, interpreting an ES value's magnitude involves combining the numeric value of an ES
124   with the context in which it was computed. Although it is not difficult to report standardized and
125   unstandardized values, the burden is placed on the research to explain their meaning given the
126   context of the study by interpreting ES magnitude (Kelley & Preacher, 2012).
127   **Meaningful Contextual Interpretations**
128   Interpreting an ES within a research context can be difficult. Asking a statistician whether
129   or not a given ES is important or meaningful, is, in all likelihood, much less useful than posing
130   the same question to a subject matter expert (Anderson, 2019). As Anderson (2019) states, what
131   implications an ES has to a field requires a thorough understanding of the utilized variables and
132   their corresponding measures, as well as the extent to which the finding reflects what is expected
133   by a theory. Other factors that can be valuable in providing meaningful interpretations include
134   how the results situate relative to studies evaluating the same hypothesis(es), the potential impact
135   of the results on individuals, families, policies, and resources (e.g., financial, labour, time)
136   required for implementing recommendations. In other words, subject matter experts are in the
137   best position to comment on extended implications of a finding in real life.
138   For example, consider a fictional example of interpreting an ES value within context in
139   basic science for research on productivity and focus. Assume a researcher is judging the effect of
140   a novel productivity exercise's effect on focus when completing a task, and finds that the
141   exercise resulted in an average increase of fifteen minutes of focus per day per employee.
142   Consider the following questions:
143   -   Is an average increase of fifteen minutes of focus per day after the exercise important?
144   -   What are the real-life implications of fifteen minutes of extra focus time?
145   -   How long are people focused on average without any intervention?
146   -   How much variability is there in the change in focus across employees?
147   -   How does the focus increase from this particular exercise compare to other productivity
148   exercises?
149   -   Is the focus task used in the exercise relevant to the actual tasks of the employees?
150   -   Is the sample used in this study representative of the population of interest?
151   -   How does the population sampled from the study translate to other populations?
152   -   Is fifteen minutes of extra focus time large enough relative to the cost of introducing the
153   productivity exercise?
154   These are just a small subset of questions that can be used to address an ES value's
155   magnitude within the context of the study; further, relevant questions can only be raised and
156   addressed by those with knowledge of the theoretical content and the finding's real world
157   implications. In other words, these questions cannot readily be answered by a statistician
158   removed from the research project, but instead should be illuminated upon by psychology
159   researchers who specialize in focus and productivity, as well as other stakeholders and experts in
160   education and workplace management.

161   It is worth noting that in this example, and extending to other areas of the paper, we are
162   interpreting ES values at the population level. However, there could be important and
163   meaningful differences at the individual level that are also worth exploring. For instance, perhaps
164   in this example, an individual who was not able to focus on any task was suddenly productive for
165   fifteen minutes per day (as opposed to zero) following the intervention. This individual ES
166   scenario might be more meaningful than increasing the productivity of a person who was focused
167   for five hours per day prior to the intervention, but focused for five hours and fifteen minutes
168   afterwards.
169   A potential critique regarding in-context interpretations of an ES value may challenge
170   whether a cumulative science can be built at all when this approach is used. Indeed, if
171   interpretations apply to one unique setting only, it is important to speak to how different ES
172   values may be combined and meta-analyzed within a broader setting. We argue that in-context
173   interpretations of an ES value's magnitude actually facilitates the building of a cumulative
174   science much better than approaches that use benchmark or cutoff values. If researchers define
175   ES magnitude using in-context interpretations, it is much easier for another researcher reviewing
176   the literature to ascertain which studies may be combined and meta-analyzed because of
177   explicitly shared contextual interpretations. Moreover, the resulting estimated effect from a
178   meta-analysis can be interpreted, and therefore expand on in-context interpretations from studies
179   addressing research questions that are identical in their intent and purpose.
180   In contrast, interpretations of an ES such as "large" and "small" are in danger of masking
181   critical contextual and study-specific information (Pek & Flora, 2018). For example, a researcher
182   combining studies which use the benchmark interpretation of an ES value may believe that they
183   are combining information that is similar because it is "context free." In actuality, it may be that
184   contextual information could reveal that studies with similar magnitudes differ greatly in terms
185   of their practical significance (highlighting again the limitations of using cut-off values).
186   Although we believe the contextual approach to be the most useful way to interpret an ES
187   value, this is often not the approach taken by researchers. The two approaches we describe below
188   are the traditional cutoff approach and the observed ES distribution-based approach.
189   **Traditional/Popular Cutoffs for ESs**
190   Cohen's d and Pearson's r are both well-known and popular ES measures among
191   psychological researchers, as they are based on well-known statistics (i.e., standardized mean
192   difference, correlation). Popular ES guidelines for Cohen's d associate a small effect with an ES
193   value that is between $d = .2$ and $d = .5$, a medium effect with an ES value that is between $d = .5$
194   and $d = .8$, and a large effect with an ES value that is greater than $d = .8$. (Cohen, 1988).
195   Similarly, popular guidelines for Pearson's r associate a small effect with an ES value that is
196   between $r = .1$ and $r = .3$, a medium effect with an ES value that is between $r = .3$ and $r = .5$, and
197   a large effect with an ES value that is greater than $r = .5$ (Cohen, 1988). While quantifying the
198   magnitude of an effect is important, ES magnitude is recommended to be interpreted within the
199   context of the specific study (Kelley & Preacher, 2012; Pek & Flora, 2018; Richard et al., 2003).

200 Accordingly, when introducing popular cutoffs in his 1988 book, Cohen states that they are
201 "recommended for use only when no better basis for estimating the index is available" (p. 25).
202 **Interpreting Effect Size Magnitude via Field-Specific Effect Size Distributions**
203   Recent methodological studies in psychology have expressed that the ES benchmark
204 values for Cohen's *d* and Pearson's *r* are unrealistic based on the observed distribution of ES
205 values in their specific field (e.g., Brydges, 2019). While some researchers convey that the
206 current threshold values are set too low (Morris & Fritz, 2013; Welsh & Knight, 2014), other
207 researchers claim that such values are set too high, making large ESs too difficult to achieve
208 (Hemphill, 2003).
209   The notion that ES benchmark values are unrealistic relative to observed ES distributions
210 in a field was first explored in a study conducted by Hemphill (2003), who considered the
211 distribution of Pearson's *r* across 380 studies that discussed psychological assessment and
212 treatment. Hemphill aimed to "extend Cohen's benchmarks by deriving empirical guidelines
213 concerning the magnitude of correlation coefficients found among psychological studies" (p. 78).
214 Notably, the distribution of observed ES values explored by Hemphill was not synonymous to a
215 theoretical probability distribution (e.g., *t* distribution). Instead, Hemphill obtained benchmark
216 values by gathering observed ES values from different studies from psychological assessment
217 and treatment, combining them into a single distribution.
218   All ES values reported were either presented as correlation coefficients or were converted
219 from Cohen's *d* to Pearson's *r* to compare observed ES value distributions. Finally, values were
220 divided equally into lower, middle, and upper thirds. Hemphill found that the lower third
221 contained ESs between *r* = 0 and *r* = .2, the middle third contained ESs between *r* = .2 and *r* = .3,
222 and the upper third contained ESs above *r* = .3. These results led Hemphill to conclude that
223 correlation values less than .2 should be considered a small effect, values between .2 and .3
224 should be considered a medium effect, and values larger than .3 should be considered a large
225 effect, to reflect such distributions.
226   Furthermore, if we convert these correlation values to a standardized mean difference
227 metric, *d* values less than .4 should be considered a small effect, values between .4 and .6 should
228 be considered a medium effect, and values larger than .6 should be considered a large effect.
229   Hemphill adopted the observed ES approach as an alternative method for interpreting ES
230 values. For instance, Hemphill noted that his findings "suggest that the value Cohen used to
231 represent a large correlation coefficient occurs somewhat infrequently in many key research
232 studies in psychology and that a lower value might be warranted in some instances (p. 79).
233 Accordingly, to interpret ES values using this approach and using Hemphill's distribution as a
234 reference, a researcher would compare their own observed ES values relative to the benchmarks
235 specified from the 280 psychological assessment and treatment studies in order to obtain a
236 "small", "medium" or "large" label (to see a demonstration of how this method works with
237 simulated data see https://mfr.ca-
238 1.osf.io/render?url=https://osf.io/34gf9/?direct%26mode=render%26action=download%26mode
239 =render.

240   This approach is extremely problematic because practical significance often becomes
241 distorted (Pek & Flora, 2018). For example, if the benchmark value for a "large" effect was set to
242 a relatively small value (such as $r = .3$ in the Hemphill study) in a specific field, there is the
243 potential for researchers to claim meaningfulness even in situations where the effect has no
244 practical significance (Pogrow, 2019). Further, working in a context-free setting,, Beribisky,
245 Davidson and Cribbie (2019) found that a correlation of $r = .3$ was the smallest association that
246 was deemed meaningful when participants viewed scatterplots of context-free associations. Thus,
247 there is simply no valid theoretical rationale for creating benchmarks of ES values for
248 researchers in psychology (or any other discipline/sub-discipline) by considering the distribution
249 of observed ES values within these disciplines (or sub-disciplines). Above all, encouraging
250 researchers to compare ES values relative to the distribution of ES within a discipline (or sub-
251 discipline) discourages the practice of interpreting ES magnitude within the context of a specific
252 study and promotes reliance on context-free benchmark values to conclude practical significance.
253 Exclusively relying on context-free benchmarks was strongly advised against by Cohen (1988) in
254 his initial writings on the topic, and countless authors since. We expand on this point below by
255 highlighting two specific problems with the field-specific observed ES distribution approach: (1)
256 equating ES magnitude with importance, and (2) publication bias.

Issue 1: Equating Effect Size Magnitude with Importance

258   One of the largest theoretical issues that exists when using the distribution of observed
259 ES values in a given field as a guideline for interpreting whether a study produces a small,
260 medium, or large effect is that the context of the study is ignored. This is of particular concern in
261 two paradoxical cases – when a study produces a numerically small ES that is considered large
262 within context, and when a study produces a numerically large ES that is considered small within
263 context.

264   Consider the following example where a statistically small ES value would be considered
265 significant given the context of the study. In 1988, a correlational analysis was conducted to seek
266 any potential relationships between the intake of aspirin and the reduced risk of death by heart
267 attack (Meyer et al., 2001). Under typical ES guidelines (e.g., Cohen, 1988), a very small effect
268 of $r = .02$ was concluded. However, researchers appropriately interpreted the ES value's
269 magnitude to be large given the context of the study for various reasons, most importantly that
270 many lives were saved. In this instance, and several others, small magnitude ES values can hold
271 large practical significance (Dunst & Hamby, 2012).

272   Similarly, consider an example in which a statistically large ES value would be
273 considered insignificant given the context of the study. An experiment conducted in the 1990s
274 analyzed the academic performance of students in high-poverty schools by implementing a
275 reading intervention program called *Success For All* (SFA). Using popular ES guidelines (e.g.,
276 Cohen, 1988), the researchers concluded a very large effect. However, the results were
277 practically insignificant because there were no meaningful improvements on the academic
278 performance of students (Pogrow, 2019). In this case, what turned out to be a poor surrogate
279 outcome could have had negative consequences (e.g., combining this ES with ESs from more

280 legitimate outcomes, potentially inflating the combined effect size. An independent study was
281 conducted a few years later and corroborated that the approach did not produce meaningful
282 improvements in performance; more specifically, the experimental group was performing poorly
283 academically relative to the control group and the program was terminated (Pogrow, 2019).
284   Thus, the main theoretical issue with the distribution of observed ES values approach is
285 that it distracts researchers from interpreting ES magnitude by considering the specific context of
286 their research study (Valentine & Cooper, 2003). In other words, the approach insinuates that
287 threshold values alone hold all of the answers about a study's practical significance. Hemphill
288 (2003) encourages the interpretation of ES magnitude within specific fields by considering the
289 distribution of ESs in that field; however, Cohen (1988), and several others (e.g., Kelley &
290 Preacher, 2018; Pek & Flora, 2018), advise us against using proposed benchmarks as a
291 determinant of practical significance alone and strongly encourage the interpretation of ES
292 magnitude by considering the specific context of the study.
293 *Issue 2: Publication Bias and Effect Sizes*
294   Publication bias can be defined as an increased likelihood for an effect that is statistically
295 significant, rather than one that is not, to be published. In psychology, statistically non-
296 significant results have sometimes been met with critiques of low power, and therefore increased
297 likelihood of Type II error (Ferguson & Heene, 2012). Due to this inconsistent treatment of
298 research findings, it has been noted that published ES values tend to be larger or inflated
299 compared to their unpublished counterparts (Vevea & Hedges, 1995). As a further complication,
300 the magnitude of publication bias is inconsistent across disciplines and even sub-disciplines of a
301 given field.
302   This creates a further difficulty for researchers using the distribution of observed ES
303 values approach when interpreting ES magnitude. Specifically, if a researcher is surveying past
304 research on a given topic or sub-discipline, they will be relying upon incomplete literature with
305 potentially over-estimated ES values. Any classification of "small", "medium", or "large" will
306 also be inflated, leading to an incorrect interpretation of ES magnitude with the distribution-
307 based approach.
308   This study contributes to the literature by defining the distribution of observed ES values
309 approach for ES interpretation and explains the problems associated with this approach. This
310 study's contributions also involve investigating whether this approach is being commonly used.
311   In the sections above, we highlighted the theoretical issues with the observed distribution
312 of ES values to interpret ES magnitude. In this section of the paper, we seek to quantify the
313 popularity of this approach by highlighting both articles recommending the approach and articles
314 adopting the approach for interpreting ES magnitude. Although the method may be theoretically
315 problematic, if researchers are ignoring it, then there is no issue. However, if this methodology is
316 gaining popularity, then there is a need to highlight the issues with the approach and provide
317 recommendations for interpreting ES magnitude.
318
319 **Survey Methodology**

320   This study is conducted in two parts. First, articles that proposed the use of field-specific
321   ES distributions to inform magnitude were identified and referred to as *primary articles*. Second,
322   in order to identify the growth of this approach, articles that cited the primary articles were
323   identified and classified by year and publication type. The first publication type consisted of
324   researchers using the primary articles to directly interpret an ES value's magnitude in their study.
325   The second type consisted of researchers using the concepts from the primary articles to interpret
326   an ES value's magnitude, but not the specific cut-offs proposed in the study. The last type
327   consisted of methodological articles.
328   **Part I: Identifying Primary Articles**
329   In addition to the Hemphill (2003) study discussed above, we identified four other studies
330   from the field of psychology that proposed ES benchmark value recommendations based on the
331   observed ES values distribution approach, namely Gignac and Szodorai (2016), Rubio-Aparicio
332   et al. (2017), Lovakov and Agadullina (2018), and Brydges (2019) (Table 1). These are
333   particularly influential because they have been published in prospering sub-fields of psychology,
334   including personality and individual differences, clinical psychology, social psychology, and
335   gerontology. Each paper also cites Hemphill's publication in an attempt to justify their
336   approaches. Further, most have been published recently, highlighting the current popularity of
337   the method.
338   Gignac & Szodorai (2016) considered the distribution of Pearson's r in personality and
339   individual differences across 87 meta-analyses from six field-specific journals. Similar to
340   Hemphill's method, all ES values were divided equally into lower, middle, and upper thirds. The
341   researchers found that the lower third contained ESs between $r = 0$ and $r = .15$, the middle third
342   contained ESs between $r = .15$ and $r = .25$, and the upper third contained ESs between $r = .25$
343   and $r = .35$. Subsequently, r values less than .15 should be considered a small effect, values
344   between .15 and .25 should be considered a medium effect, and values between .25 and .35
345   should be considered a large effect. The authors concluded that differential psychologists will
346   "be arguably better served by applying the correlation guidelines reported in this investigation
347   rather than those reported by Cohen (1988, 1992) or even Hemphill (2003), in order to obtain an
348   accurate sense of the magnitude" (Gignac & Szodorai, 2016).
349   Rubio-Aparicio and colleagues (2017) considered the distribution of Cohen's d in clinical
350   psychology from 54 meta-analyses. The authors recommended benchmark values for researchers
351   in clinical psychology such that between $d = 0$ and $d = .249$ should be considered a small effect,
352   between $d = .249$ and $d = .409$ should be considered a medium effect, and between $d = .409$ and
353    $d = .695$ should be considered a large effect (Rubio-Aparicio et al., 2017. They also state that
354   "this classification should be used instead of Cohen's (1988) proposal, for the interpretation of
355   the standardized mean change values in the clinical psychological context" (Rubio-Aparicio et
356   al., 2017).
357   Lovakov & Agadullina (2018) gathered 161 meta-analyses among 29 social psychology
358   journals and reported the lower, middle, and upper thirds of Pearson's r. They recommended ES
359   values between $r = 0$ and $r = .1$ to be considered a small effect, between $r = .1$ and $r = .25$ to be

360   considered a medium effect, and between $r = .25$ and $r = .4$ to be considered a large effect
361   (Lovakov & Agadullina, 2018), highlighting that Cohen's (1988) guidelines were too
362   conservative.
363   Most recently, Brydges (2019) focused on the distribution of Pearson's r in gerontology
364   by collecting 88 meta-analyses from 10 field-specific journals. He proposed ES values less than r
365     $= .1$ to be considered a small effect, between $r = .1$ and $r = .2$ to be considered a medium effect,
366   and between $r = .2$ and $r = .3$ to be considered a large effect (Brydges, 2019). Again, Brydges
367   (2019) discusses how Cohen's (1988) ES benchmark values were set too high. Similar to the
368   other authors, Brydges (2019) also ignores the fact that Cohen (1988) suggested only using such
369   guidelines when no contextual information is available.
370   **Part II: Citations of Primary Articles**
371   In order to determine if the recommendations of the primary articles found in Part I were
372   being adopted by researchers, we coded articles that cited Hemphill (2003), Gignac & Szodorai
373   (2016), or Rubio-Aparicio et al. (2017) as found on *Google Scholar* (2020). The research
374   conducted by Lovakov & Agadullina (2018) and Brydges (2019) were excluded from the coding
375   procedure, as they were published too recently to observe any trends within the citations.
376   First, the team met to derive a coding checklist and discuss the criteria for selecting
377   specific classifications. In order to evaluate the reason behind the citation of one of the three
378   selected primary articles (i.e., to verify that it was to interpret an ES value using field-specific
379   observed ES distributions as opposed to another reason), six coding categories were developed.
380   Publications that used the distribution of observed ES values approach proposed by Hemphill
381   (2003), Gignac & Szodorai (2016), or Rubio-Aparicio (2017) to interpret ES magnitude were
382   coded as *Direct Interpretations*; papers that used the observed ES value distribution concept
383   from Hemphill (2003), Gignac & Szodorai (2016), or Rubio-Aparicio (2017), but different
384   benchmark values to interpret ES magnitude were coded as *Conceptual Interpretations*; papers
385   that were of a methodological or theoretical nature were coded as *Methodological/Theoretical*;
386   and lastly papers that cited a primary article for a reason unrelated to the interpretation of an ES
387   were coded as *Irrelevant*. Articles that used more than one method for interpreting ES magnitude
388   (e.g., Hemphill's (2003) approach and Cohen's (1988) cutoffs) were coded as *Multiple Methods,*
389   and articles that could not be accessed were coded as *Inaccessible.* Articles that used a direct or
390   conceptual interpretation of ES values demonstrate the popularity of applying the field-specific
391   observed ES distributions method to interpret ES values.
392   Books, dissertations, theses, and non-English articles with no available translations were
393   excluded from the sample.
394   To establish interrater reliability, the team (the three authors plus five research assistants)
395   used the checklist to code ten articles selected at random that cited any of the three primary
396   articles. The group met to discuss their findings and found a total of 15 discrepancies across 320
397   items in total (10 articles x 8 coders x 4 items), offering 95.3% agreement among coders. The
398   discrepancies were discussed in order to clarify any issues regarding the coding. Further, a
399   decision regarding any controversial categorizations was made by the authors.

400

Results

401

402   Overall, 426 articles cited Hemphill (2003), 122 articles cited Gignac & Szodorai (2015),
403 and 10 articles cited Rubio-Aparicio et al. (2017), for a total of N = 686 citations. A steady
404 increase in the number of citations to each primary article from peer-reviewed research articles
405 between one year following the publication date of the primary article until 2019 can be
406 observed in Figure 1. Amongst these peer-reviewed articles, 323 studies (62.5%) were coded as
407 *Direct Interpretations*, 10 articles (1.9%) were coded as *Conceptual Interpretations*, 35 articles
408 (6.8%) were coded as *Theoretical /Methodological*, and 37 articles (7.2%) were coded as
409 *Irrelevant*. As well, 90 of the above articles (17.4%) were also coded as *Multiple Methods*, all of
410 which were initially coded as a *Direct Interpretation*. Finally, 21 articles (4.3%) were coded as
411 *Inaccessible* (Figure 2).

412

Discussion

413

414   Popularizing the use of the distribution of observed ES values within a field weakens the
415 practice of interpreting an ES value's magnitude within the context of a study by promoting
416 reliance on ES benchmark values, as advised against by Cohen (1988) in his initial proposals.
417 Hemphill (2003), Gignac and Szodorai (2016), Rubio-Aparicio et al. (2017), Lovakov and
418 Agadullina (2018), and Brydges (2019) have supported the distribution of observed ES values
419 approach by proposing field-specific ES benchmarks, which are becoming increasingly
420 influential in psychological assessment and treatment, personality and individual differences, and
421 clinical psychology, subsequently. Our review found that citations to papers proposing field-
422 specific ES benchmarks are growing rapidly within sub-disciplines of psychology as researchers
423 seek suggestions for interpreting the magnitude of ES values. Furthermore, most peer-reviewed
424 articles that cited these primary articles used the field-specific ES benchmarks in the papers to
425 interpret ES magnitude; i.e., they adopted the distribution of observed ES values approach for
426 informing magnitude relative to their own obtained ES (rather than, for example, just mentioning
427 the method). This is extremely problematic, as there is little that can be gained about the
428 meaningfulness of a particular effect, within a particular context, by exploring the distribution of
429 ES values within the narrowly or broadly defined discipline. Moreover, these alternative methods
430 for interpreting ES magnitude discourages the practice of interpreting ES magnitude within the
431 context of the study itself.
432   To ensure that the observed increase in the number of primary article citations was not a
433 result of an increasing number of journals in psychology overall, we conducted a literature
434 search through *Google Scholar* (2020) to identify the total number of published articles in
435 journals containing the keyword "psychology" between 2003 and 2019. Interestingly, the total
436 number of articles published in journals containing the keyword "psychology" has remained
437 fairly stable over this time frame (Figure 1). Thus, the increase in the number of citations
438 received by Hemphill (2003), Gignac and Szodorai (2016), Rubio-Aparicio et al. (2017),
439 Lovakov and Agadullina (2018), and Brydges (2019) do not appear to reflect any overall

440 increase in the number of articles published within the discipline, discounting the reference
441 group problem.
442   To be clear, we are not arguing that the distribution of ES values in a given field or
443 discipline is of zero value; it is very important to understand the magnitude of ES values in a
444 given discipline and to appreciate how ES magnitude varies within, and across, disciplines. Our
445 point is that researchers should not interpret the magnitude of a particular effect, which occurs
446 within a specific context, by using cut-offs that are derived from an exploration of the
447 distribution of ES values in a broad discipline. For example, when interpreting the magnitude of
448 the relationship between aspirin and heart attacks (Meyer et al., 2001), the distribution of ES
449 values across the discipline at large are trivial compared to the implications of the given study
450 (where the only contextually shared information might be the overarching discipline itself).
451   One limitation of this study is that we conducted the search in *Google Scholar*. Although
452 we expect the trend of citations to generalize to other search engines and databases, our findings
453 concern *Google Scholar* results only.
454

Conclusions

456   Using the distribution of observed ES values approach for interpreting ES magnitude
457 distracts researchers from informing magnitude by considering the specific context of their
458 research study. The upwards trend in citations received by the primary articles (e.g., Hemphill,
459 2003) demonstrates an increase in the attention that this malpractice is earning, and the fact that
460 most citations are for direct interpretations of ES magnitude confirms our hypothesis that
461 researchers are adopting the approach as a crutch for informing magnitude. We hope that this
462 study is valuable in highlighting to researchers both the theoretical issues with the distribution of
463 observed ES values approach, as well as the rising popularity of the strategy; a dangerous
464 combination. Researchers in psychology lack clear guidelines regarding the process of
465 interpreting ES magnitude (e.g., Farmus et al., 2020; Funder & Ozer, 2019), leaving them
466 susceptible to adopting practices that may be of limited value. Psychology is in desperate need of
467 more research regarding best practices for interpreting ES magnitude, but while we wait for these
468 recommendations, journal editors need to encourage researchers to interpret ES magnitude using
469 any contextual information available. Further, it is important that textbook authors, statistics
470 instructors, and the like highlight the importance of not using published benchmark values for
471 interpreting ES magnitude and instead encourage researchers to consider contextual factors.
472 Improving ES magnitude interpretation will not occur overnight, but it is with hope that these
473 recommendations will encourage researchers to consider what factors are valuable when
474 considering the practical significance of their results.
475

Acknowledgements

477   We would like to thank our research assistants Farwa Abdi, David Civil, David Eldridge,
478 Korina Taguba, and Anh Truong Phuong Nguyen for assisting in the coding process.
479

References

480 **References**

481 Anderson, A. A. (2019). Assessing statistical results: Magnitude, precision, and model

482   uncertainty. *American Statistician, 73*(sup1), 118-121.

483   https://doi.org/10.1080/00031305.2018.1537889

484 Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British*

485   *Journal of Psychology, 100*(3), 603-617. https://doi.org/10.1348/000712608X377117

486 Beribisky, N., Davidson, H., & Cribbie R.A. (2019). Exploring perceptions of meaningfulness

487   in visual representations of bivariate relationships. *PeerJ, 7*.

488   https://doi.org/ 10.7717/peerj.6853

489 Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect

490   size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449.

491   https://doi.org/10.1037/a0038047

492 Brydges, C. R. (2019). Effect Size Guidelines, Sample Size Calculations, and Statistical

493   Power in Gerontology. *Innovation in Aging*, *3*(4). https://doi.org/10.1093/geroni/igz036

494 Burchinal, M. R. (2008). How measurement error affects the interpretation and understanding of

495   effect sizes. *Child Development Perspectives, 2*(3), 178-180.

496   https://doi.org/10.1111/j.1750-8606.2008.00062.x

497 Bushman, B. J., Rothstein, H. R., & Anderson, C. A. (2010). Much ado about something:

498   Violent video game effects and a school of red herring: Reply to Ferguson and Kilburn

499   (2010). *Psychological Bulletin*, *136*(2), 182–187. https://doi.org/10.1037/a0018718

500 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L.

501   Erlbaum Associates.

502 Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

503   https://doi.org/10.1037/0033-2909.112.1.155

504 Cooper, H., & Findley, M. (1982). Expected effect sizes. *Personality and Social Psychology*

505   *Bulletin,8*(1), 168-173. doi:10.1177/014616728281026

506 Dunst, C. J., & Hamby, D. W. (2012). Guide for calculating and interpreting effect sizes and

507   confidence intervals in intellectual and developmental disability research studies.

508   *Journal of Intellectual & Developmental Disability*, *37*(2), 89–99.

509   https://doi.org/10.3109/13668250.2012.673575

510 Gignac, G. E., & Szodorai, E. T. (2015). Effect size guidelines for individual differences

511   researchers. *Personality and Individual Differences*, *102*, 74–78.

512   https://doi.org/10.1016/j.paid.2016.06.069

513 Farmus, L., Beribisky, N., Gutierrez, N. M., Alter, U., Panzarella, E., & Cribbie, R. (2020,

514   November 9). Effect size reporting and interpretation in social personality research.

515   https://doi.org/10.31234/osf.io/nvczj

516 Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and

517   psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6),

518   555-561. https://doi.org/10.1177/1745691612459059

519   Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality
520   research: Current practice and recommendations. *Social Psychological and Personality*
521   *Science, 8*(4), 370-378. https://doi.org/10.1177/1948550617693063
522   Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense
523   and nonsense. *Advances in Methods and Practices in Psychological Science,2*(2),
524   156-168. http://doi.org/10.1177/2515245919847202
525   Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences
526   researchers. *Personality and Individual Differences*, *102*, 74–78.
527   http://doi.org/10.1016/j.paid.2016.06.069
528   Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related
529   estimators. *Journal of Educational Statistics*, *6*(2), 107–128.
530   https://doi.org/10.3102/10769986006002107
531   Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American*
532   *Psychologist*, *58*(1), 78 -79. https://doi.org/10.1037/0003-066x.58.1.78
533   Kelley, K., & Preacher, K. J. (2011). Effect size measures for mediation models: Quantitative
534   strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93–115.
535   http://doi.org/10.1037/a0022658
536   Lovakov, A., & Agadullina, E. (2017). Empirically Derived Guidelines for Interpreting Effect
537   Size in Social Psychology. https://doi.org/10.1002/ejsp.2752
538   Manolov, R., Jamieson, M., Evans, J. J., & Sierra, V. (2016). A discussion of alternatives for
539   establishing empirical benchmarks for interpreting single-case effect sizes.
540   *Psicológica*, *37*, 209–234.
541   Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G.
542   M. (2001). Psychological testing and psychological assessment: A review of evidence
543   and issues. *American Psychologist,56*(2), 128-165. https://doi.org/10.1037/0003-
544   066x.56.2.128
545   Morris, P. E., & Fritz, C. O. (2013). Effect sizes in memory research. *Memory,21*(7), 832-842.
546   http://doi.org/10.1080/09658211.2013.763984
547   Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical
548   significance: a practical guide for biologists. *Biological Reviews, 82*(4), 591-605.
549   https://doi.org/10.1111/j.1469-185X.2007.00027.x
550   Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A
551   discussion and tutorial. *Psychological Methods*, *23*(2), 208–225.
552   https://doi.org/10.1037/met0000126
553   Pogrow, S. (2019). How Effect Size (Practical Significance) Misleads Clinical Practice: The
554   Case for Switching to Practical Benefit to Assess Applied Research Findings. *The*
555   *American Statistician*, *73*(sup1), 223–234.
556   https://doi.org/10.1080/00031305.2018.1549101
557   Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative
558   strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93–115.

559   https://doi.org/10.1037/a0022658

560   Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social
561   Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363.
562   https://doi.org/10.1037/1089-2680.7.4.331

563   Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2018). A
564   methodological review of meta-analyses of the effectiveness of clinical psychology
565   treatments. *Behavior Research Methods*, *50*(5), 2057–2073.
566   https://doi.org/10.3758/s13428-017-0973-8

567   Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.),
568   *The Handbook of Research Synthesis* (pp. 231-244). Sage.

569   Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological
570   Research: Differences Between Sub-Disciplines and the Impact of Potential Biases.
571   *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813

572   Seligman, L. D., Ollendick, T. H., Langley, A. K., & Baldacci, H. B. (2004). The Utility of
573   Measures of Child and Adolescent Anxiety: A Meta-Analytic Review of the Revised
574   Children's Manifest Anxiety Scale, the State–Trait Anxiety Inventory for Children, and
575   the Child Behavior Checklist. *Journal of Clinical Child & Adolescent Psychology*,
576   *33*(3), 557–565. https://doi.org/10.1207/s15374424jccp3303_13

577   Simpson, A. (2018). Princesses are bigger than elephants: Effect size as a category error in
578   evidence-based education. *British Educational Research Journal*, *44*(5), 897–913.
579   https://doi.org/10.1002/berj.3474

580   Valentine, J. C. & Cooper, H. (2003). Effect size substantive interpretation guidelines:
581   Issues in the interpretation of effect sizes. Washington, DC: What Works
582   Clearinghouse.

583   Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the
584   presence of publication bias. *Psychometrika*, *60*(3), 419-435.
585   https://doi.org/10.1007/BF02294384

586   Welsh, A. H., & Knight, E. J. (2014). ''Magnitude-based inference'': A statistical review.
587   *American College of Sports Medicine,* 874-884.
588   http://doi.org/10.1249/MSS.0000000000000451

Figure 1

The number of citations each primary article has received since their publication date versus the number of articles published in a journal containing the keyword 'psychology'.

This histogram represents the number of citations received from the following articles: Hemphill (2003), Gignac & Szodorai (2016), Rubio-Aparicio et al. (2017), Lovakov & Agadullina (2018), and Brydges (2019). The line represents the number of articles that were published in a journal containing the keyword 'psychology'.

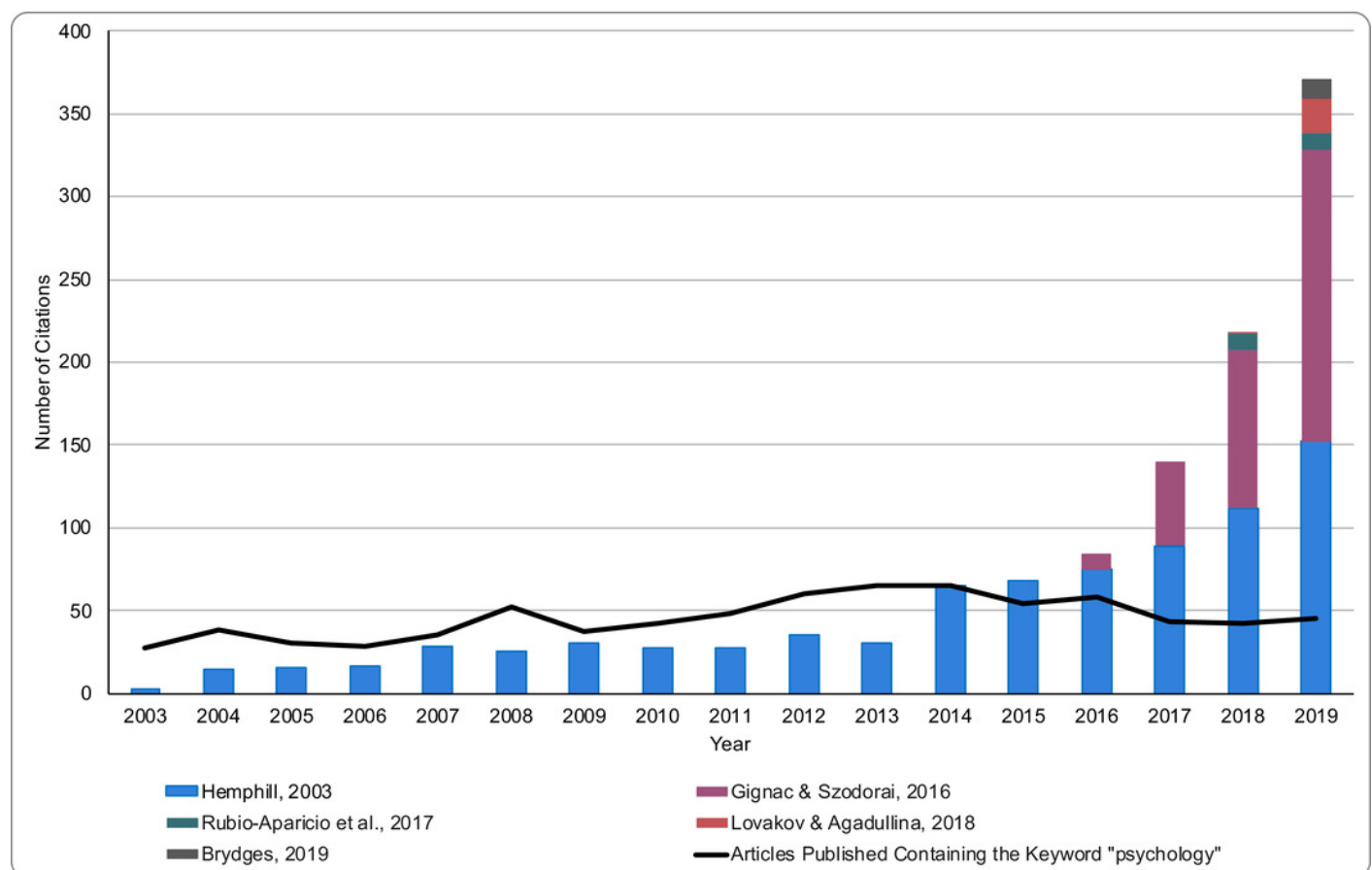
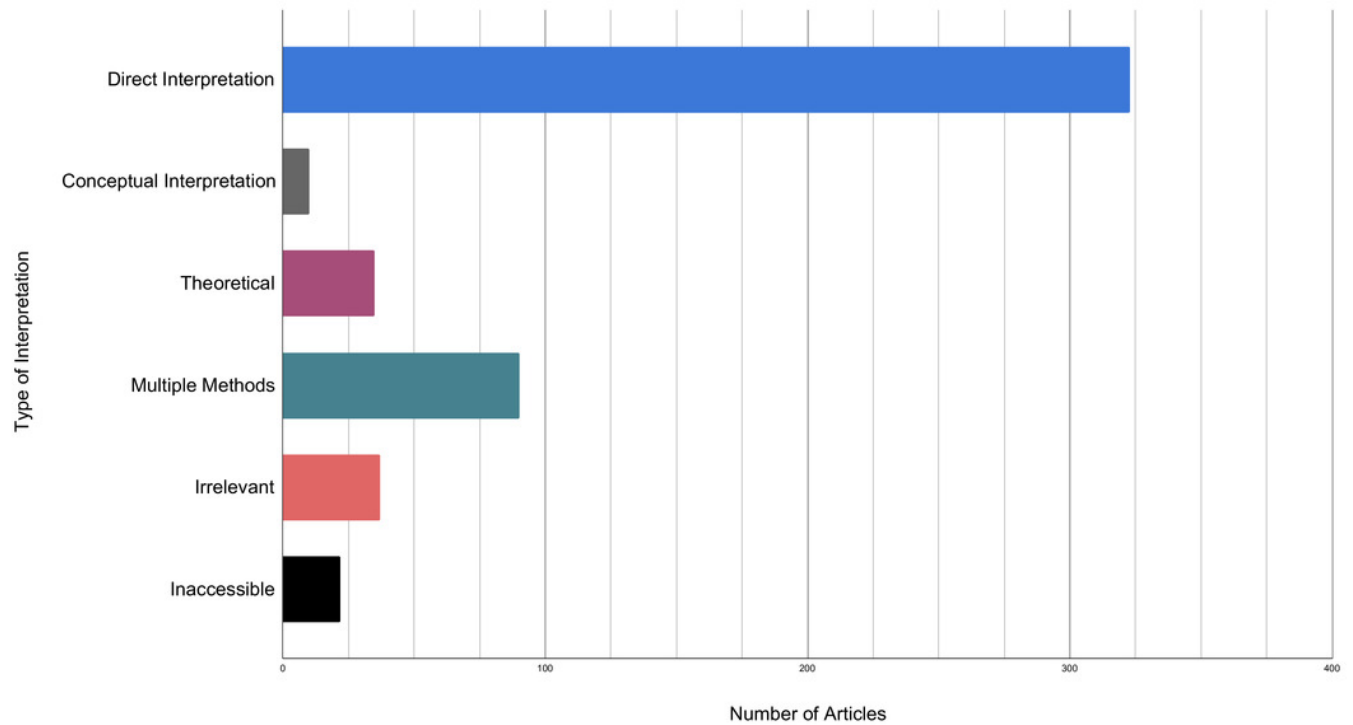



Figure 2

The types of citations received by Hemphill (2003), Gignac & Szodorai (2016), and Rubio-Aparicio et al. (2017).

Articles were coded on the basis of: (1) direct interpretation; papers that used the observed distribution of ES values from a primary article to inform magnitude, (2), conceptual interpretation; papers that used the concept of observed distribution of ES values, but different benchmark values to inform magnitude; (3) methodological / theoretical, (4) irrelevant; papers that cited a primary article for a reason unrelated to the interpretation of an ES value, (5) multiple methods; papers that used more than one method for interpreting ES magnitude, and (6) inaccessible; articles that could not be accessed.

Field-specific effect size (ES) benchmarks proposed by primary articles to interpret ES magnitude.

Five primary articles propose field-specific ES benchmarks by considering the distribution of Pearson's *r* or Cohen's *d* in their field. Values of *r* or *d* were divided into lower, middle, and upper thirds to create these benchmarks.

Field of Study	Type of ES	Proposed ES Benchmark Values		
		Small	Medium	Large
Original Benchmark Values **Cohen (1988)**	*d*	0.2	0.5	0.8
	r	0.1	0.3	0.5
Psychological Assessment and Treatment **Hemphill (2003)**	*r*	< 0.2	0.2	0.3
Personality and Individual Differences **Gignac & Szodorai (2015)**	*r*	0.15	0.25	0.35
	d	0.1	0.3	0.5
Clinical Psychology **Rubio-Aparicio et al. (2017)**	*d*	0.249	0.409	0.695
Social Psychology **Lovakov & Agadullina (2018)**	*r*	0.1	0.25	0.4
Gerontology- Psychology of Aging **Brydges (2019)**	*r*	0.1	0.2	0.3