International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Is Our Result Unusual? A Practical Procedure for Benchmarking Small-Sample Correlations against Meta-Analytic Evidence

Issah Quarmson

Doctor of Business Administration Programme, Westford University College and Universidad Católica San Antonio de Murcia (UCAM)

Abstract

Practitioner researchers, including doctoral candidates in professional programmes, often estimate correlations from samples of around fifty and compare them with meta-analytic estimates. The comparison is usually informal: a result is called consistent with, or weaker than, the literature without asking whether the gap exceeds what sampling error, measurement error and genuine between-study variation would produce. This paper sets out a transparent six-step procedure for making the comparison properly: select a commensurable benchmark; estimate a confidence interval for the observed correlation using Fisher's z transformation; attenuate the corrected benchmark to the reliabilities of the study's own measures; compare the observed correlation with both the benchmark mean and the distribution of true effects implied by the meta-analysis; stress-test the conclusion; and calculate what the study could have detected. The procedure is illustrated with six correlations from a survey of 52 expatriate educators in Guangzhou, China, benchmarked against Tett and Meyer (1993) and Meyer et al. (2002). None of the six correlations differed reliably from its benchmark. An apparent shortfall in the satisfaction and turnover intention correlation ($r = -0.26$ against a published estimate of -0.58) was significant ($p = 0.046$) under only one combination of analytic choices. It was no longer significant with a construct-matched benchmark ($p = 0.107$), with allowance for between-study heterogeneity ($p = 0.167$), after correction for multiple testing, with a bootstrap interval, or when a single influential respondent was excluded. With 52 respondents, only differences of roughly 0.20 to 0.40 correlation units were detectable with 80 percent power. Researchers should choose benchmarks that match their constructs and measures, put both correlations on the same metric, distinguish a result that differs from the average study from one that is unusual among studies, and report what their sample could and could not have detected. Spreadsheet formulas are provided for every step.

Keywords: Meta-Analysis, Benchmarking, Correlation, Fisher z Transformation, Credibility Interval, Prediction Interval.

1. Introduction

Professional doctorates and applied research projects often rely on surveys of modest size. A researcher studying teachers, nurses or managers in one city or sector may reasonably obtain 50 to 100 usable responses. Having estimated the correlations among the variables of interest, the researcher naturally turns to the literature for context. For many relationships in organisational research, meta-analyses

summarise evidence from tens of thousands of respondents, and their estimates serve as natural benchmarks [1].

The comparison is usually handled informally. A correlation of −0.26 is described as "weaker than the meta-analytic estimate of −0.58", and the discussion turns to why the local context might weaken the relationship. A correlation of 0.73 is described as "consistent with" a benchmark of 0.70. Neither statement is necessarily wrong, but neither has been tested. The first gap may reflect nothing more than sampling error. The second apparent agreement may hide a difference in the opposite direction, because the benchmark and the observed correlation are not on the same metric.

General guidance on interpreting effect sizes is plentiful. The conventions of Cohen (1988) [2], and the empirically grounded alternatives of Bosco et al. (2015) [1] and Gignac and Szodorai (2016) [3], help researchers judge whether a correlation is small or large relative to a field. It has also been stressed that point estimates from small samples are unstable [4] and should be reported with intervals [5]. What is missing is a compact, practical procedure for the specific task of comparing one small-sample correlation with one meta-analytic estimate, a task that is ubiquitous in theses and applied articles yet rarely carried out formally.

This paper offers such a procedure and makes three contributions. First, it assembles established techniques, namely Fisher's transformation [6, 7], the classical correction for attenuation [8, 9] and power analysis, into a six-step workflow that requires nothing more than a spreadsheet. Second, it distinguishes two questions that informal comparisons conflate: whether a result differs from the average study, and whether it is unusual given the genuine variation among studies that meta-analyses routinely report [10, 11]. Third, a worked example shows that analytic choices, above all the choice of benchmark, can matter more than the data, and demonstrates how that dependence can be reported honestly.

Section 2 explains why informal comparisons mislead. Section 3 sets out the procedure, and Section 4 applies it to six correlations from a survey of expatriate educators. Section 5 discusses the lessons and offers reporting recommendations, Section 6 notes limitations, and Section 7 concludes.

2. Why Informal Comparisons Mislead

2.1 Sampling Error Is Large

With 52 respondents the standard error of Fisher's z is $1/\sqrt{49} \approx 0.14$, so a 95% confidence interval for a correlation of moderate size spans roughly 0.4 to 0.5 correlation units: about 0.50 units when r is near ±0.30 and about 0.42 when r is near ±0.50. The interval around r = −0.26, for example, runs from −0.50 to +0.01. A single estimate of this kind is compatible with a small, medium or large population correlation by the conventions of Cohen (1988) [2], and many apparent differences from the literature fall comfortably inside it.

2.2 Benchmarks Are Usually on a Different Metric

Most meta-analyses in organisational research follow the psychometric approach, in which each observed correlation is corrected for measurement unreliability before averaging [9]. The resulting estimate, ρ, is the correlation expected if both variables were measured without error. A single study's observed correlation is not corrected. Because correction always increases the magnitude of a correlation, comparing an observed r directly with ρ is biased towards concluding that the observed relationship is weaker, and towards masking cases in which it is stronger.

2.3 A Benchmark Is a Distribution, Not a Constant

Meta-analyses estimate not only the mean true correlation but also its standard deviation across settings,

SD_ρ, often summarised as a credibility interval [10]. For the relationships examined here, SD_ρ ranges from about 0.08 to 0.25 (Table 1). When true correlations vary this much across organisations, occupations and countries, a local study can differ from the average for substantive reasons and still be entirely ordinary. Asking whether a result differs from the mean is therefore a different question from asking whether it is unusual [11, 12].

2.4 Constructs and Measures Must Match

Meta-analytic estimates are specific to the operationalisations they pool. Tett and Meyer (1993) [13], for instance, show that job satisfaction correlates more strongly with broad withdrawal-cognition scales than with narrowly worded turnover-intention scales, and their commitment estimates refer to affective commitment only. A benchmark drawn from a different construct or measure type can create, or conceal, a difference that has nothing to do with the study's context.

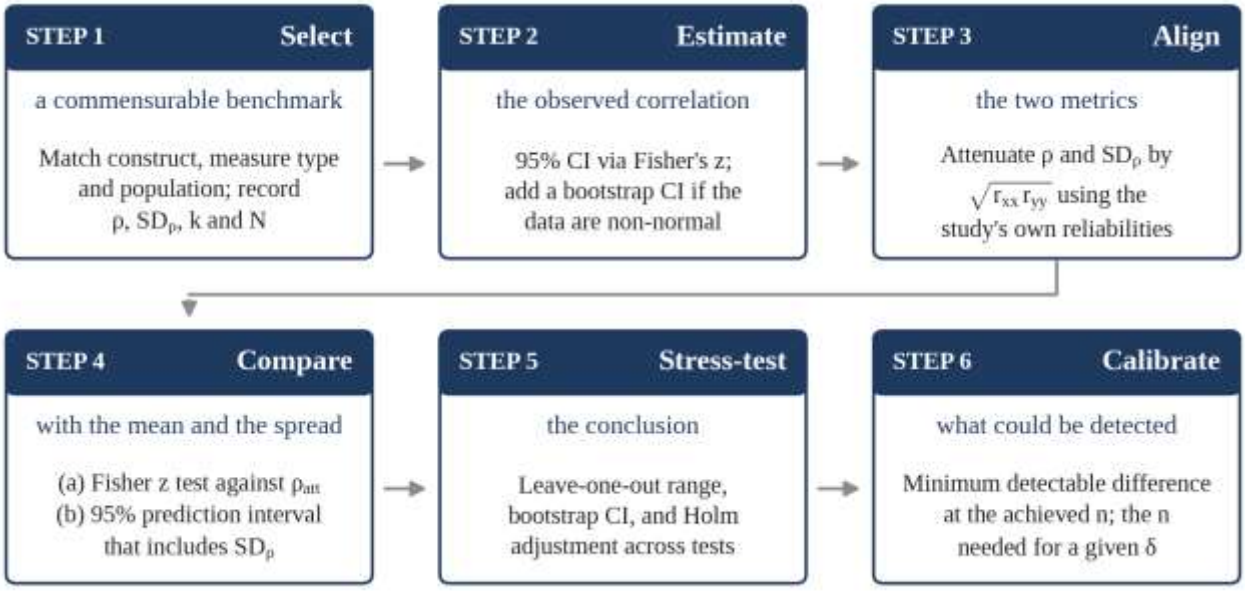
2.5 Small Samples Are Fragile, and Comparisons Are Multiple

A single unusual respondent can move a correlation substantially when n = 52, and researchers who benchmark several correlations at once face the familiar inflation of false positives across tests [14]. A difference that disappears when one case is removed, or that would not survive adjustment for the number of comparisons made, should be reported as tentative.

3. The Procedure

Figure 1 summarises the six steps. Equations are numbered for reference, and the equivalent spreadsheet formulas are given in Section 8.2.

Figure 1: Six-Step Procedure for Benchmarking a Small-Sample Correlation against a Meta-Analytic Estimate



Note: Steps 1 to 4 produce the comparison; Steps 5 and 6 qualify how much weight it can bear.

3.1 Step 1 – Select a Commensurable Benchmark

A meta-analytic estimate should be identified whose constructs, measure types and, ideally, population resemble those of the study. Many meta-analyses report moderator subgroups (for example, multi-item versus single-item measures, or North American versus other samples). The subgroup closest to the study's design is usually the better benchmark, even if it rests on fewer studies. The corrected mean ρ, the true-score standard deviation SD_ρ (or the credibility interval from which it can be recovered), the number of studies k and the total sample N should be recorded. Where no close match exists, this should be stated and the comparison treated as indicative only.

3.2 Step 2 – Estimate a Confidence Interval for the Observed Correlation

Because the sampling distribution of r is skewed when the population value is far from zero, the interval is computed on the z scale of Fisher (1915) [6]:

$$z_r = \operatorname{artanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \quad SE(z_r) = \frac{1}{\sqrt{n-3}} \tag{1}$$

$$95\%\ \text{CI} = \tanh\left[z_r \pm \frac{1.96}{\sqrt{n-3}}\right] \tag{2}$$

In a spreadsheet, the functions FISHER and FISHERINV perform the transformations. Fisher's interval assumes approximate bivariate normality. Where the data are skewed or contain outliers, a percentile bootstrap interval [15] should be reported alongside it, because Fisher intervals can be too narrow under non-normality [16].

3.3 Step 3 – Put the Benchmark on the Study's Metric

If the benchmark is corrected for unreliability, both its mean and its standard deviation should be attenuated to the reliabilities of the study's own measures [8]:

$$\rho_{att} = \rho \times \sqrt{r_{xx} \times r_{yy}}, \quad SD_{att} = SD_\rho \times \sqrt{r_{xx} \times r_{yy}} \tag{3}$$

Here r_{xx} and r_{yy} are the study's reliability estimates, typically coefficient alpha [17]. The alternative, disattenuating the observed correlation, is less attractive: it inflates the observed correlation's sampling error along with the estimate and can produce values beyond ±1 in small samples. Attenuating the benchmark leaves the observed correlation and its interval untouched.

3.4 Step 4 – Compare the Observed Correlation with the Mean and with the Distribution

Two questions should be kept separate. The first is whether the study differs from the average study. Treating ρ_{att} as a fixed value, which is reasonable when the meta-analytic mean rests on thousands of respondents, the difference is tested on the Fisher scale:

$$z_M = [\operatorname{artanh}(r) - \operatorname{artanh}(\rho_{att})] \times \sqrt{n-3} \tag{4}$$

Equivalently, the difference is significant at the 5% level when ρ_{att} lies outside the 95% interval from Step 2. The second question is whether the result is unusual given genuine variation among settings. If true correlations vary with standard deviation SD_{att}, an observed correlation from a new study of size n is expected to fall within a prediction interval that combines that variation with sampling error [9, 11]:

$$z_H = \frac{r - \rho_{att}}{\sqrt{SD_{att}^2 + \frac{(1-\rho_{att}^2)^2}{n-1}}} \tag{5}$$

$$95\%\ \text{PI} = \rho_{att} \pm 1.96 \times \sqrt{SD_{att}^2 + \frac{(1-\rho_{att}^2)^2}{n-1}} \tag{6}$$

A result inside this interval is not unusual: it is the kind of value that a study of this size would often produce somewhere in the population of settings the meta-analysis represents. The test against the mean is the more sensitive of the two, but only the comparison with the distribution addresses whether a result is unusual.

3.5 Step 5 – Stress-Test the Conclusion

Three checks are recommended for small samples. A leave-one-out analysis recomputes the correlation n times, omitting one case each time, to show how far a single respondent can move it. A bootstrap interval checks whether the Fisher interval is too narrow. When several correlations are benchmarked, p values should be adjusted for multiplicity; the sequential procedure of Holm (1979) [14] controls the familywise error rate and is uniformly more powerful than the Bonferroni correction. An influential case should be reported together with the results with and without it, rather than silently removed.

3.6 Step 6 – Calibrate What Could Have Been Detected

A non-significant difference means little unless the study could have detected a difference of meaningful size. The sample needed to detect a true difference δ from ρ_{att} with power $1 - \beta$ at two-sided level α is:

$$n = \left[\frac{z_{1-\alpha/2} + z_{1-\beta}}{|\operatorname{artanh}(\rho_{att} + \delta) - \operatorname{artanh}(\rho_{att})|}\right]^2 + 3 \quad (7)$$

For the conventional choices, $z_{0.975} + z_{0.80} \approx 2.80$. Conversely, for the achieved n, the correlations that could just be distinguished from the benchmark in each direction are:

$$r_{detectable} = \tanh\left[\operatorname{artanh}(\rho_{att}) \pm \frac{2.80}{\sqrt{n-3}}\right] \quad (8)$$

Reporting these quantities makes explicit that "no significant difference" from a small study usually means "no evidence of a large difference", not "no difference" [18].

4. Worked Example

4.1 Study and Measures

The example uses a survey of 52 expatriate educators in international high schools in Guangzhou, China [19]. Job satisfaction was measured with 12 items ($\alpha = 0.89$). Organisational commitment was measured with nine items covering the affective, continuance and normative components of the model of Meyer and Allen (1991) [20], three items per component (affective $\alpha = 0.89$, continuance $\alpha = 0.61$, normative $\alpha = 0.86$; nine-item composite $\alpha = 0.85$). Turnover intention was measured with three items ($\alpha = 0.84$). Six correlations were benchmarked: satisfaction with overall commitment and with turnover intention; overall commitment with turnover intention; and each commitment component with turnover intention. All computations below use correlations rounded to two decimals, as reported in the parent study, so values computed from the raw data may differ in the third decimal.

4.2 Selecting the Benchmarks

Table 1 lists the benchmarks and rates how well each matches the study (Step 1). Two choices deserve comment.

First, the headline estimate of Tett and Meyer (1993) [13] for satisfaction and turnover intention, $\rho = -0.58$, pools narrowly worded turnover-intention scales with broader withdrawal-cognition scales, which also ask about thinking of quitting and searching for alternative work. Their subgroup analysis shows

that the two families behave differently: for multi-item turnover-intention scales the estimate is ρ = −0.53 (k = 16), whereas for withdrawal-cognition scales it is ρ = −0.74 (k = 26). Because the study's three-item scale measures intention to leave, the turnover-intention subgroup is the commensurable benchmark. The same subgroup gives ρ = −0.52 for commitment and turnover intention, in place of the pooled −0.54.

Second, the commitment estimates of Tett and Meyer (1993) [13] refer to affective commitment only, measured mainly with the Organizational Commitment Questionnaire. The study's commitment composite also contains continuance and normative items, so the two composite-based comparisons are only partial matches and are retained for illustration. The component-level estimates of Meyer et al. (2002) [21], based on the Affective, Continuance and Normative Commitment Scales [22], are close matches for the three component correlations, although their outcome category, withdrawal cognition, is somewhat broader than turnover intention.

Table 1 also shows how heterogeneous the benchmarks are. The 90% credibility interval for affective commitment and withdrawal cognition runs from −0.90 to −0.21, and that for normative commitment crosses zero. Meyer et al. (2002) [21] further report that the affective commitment relationship is weaker in studies outside North America (ρ = −0.49) than within it (ρ = −0.58), a reminder that the population a benchmark represents is itself a modelling choice.

Table 1: Meta-Analytic Benchmarks and Their Correspondence with the Study's Measures

Relationship	Source and Subgroup	k	N	ρ	SD_ρ	90% CrI	Match
Satisfaction–commitment	Tett and Meyer (1993), Table 3: multi-item facet satisfaction	36	25,340	0.70	0.08	[0.56, 0.84]	Partial[a]
Satisfaction–intention	Tett and Meyer (1993), Table 5: multi-item turnover intention	16	3,602	−0.53	0.11	[−0.71, −0.35]	Close
Commitment–intention	Tett and Meyer (1993), Table 5: multi-item turnover intention	9	2,392	−0.52	0.14	[−0.75, −0.29]	Partial[a]
Affective–intention	Meyer et al. (2002): overall withdrawal cognition	51	17,282	−0.56	0.21	[−0.90, −0.21]	Close[b]
Normative–intention	Meyer et al. (2002): overall withdrawal cognition	25	9,645	−0.33	0.25	[−0.74, 0.08]	Close[b]
Continuance–intention	Meyer et al. (2002): overall withdrawal cognition	39	13,590	−0.18	0.15	[−0.42, 0.07]	Close[b]

Note: ρ = mean correlation corrected for unreliability; SD_ρ = estimated standard deviation of true correlations; CrI = credibility interval, ρ ± 1.645 SD_ρ. For Tett and Meyer (1993) [13], SD_ρ is the square root of the reported residual variance of corrected correlations; credibility intervals for Meyer et al. (2002) [21] are as published. [a] Benchmark commitment is affective commitment, whereas the study uses a three-component composite. [b] Benchmark outcome (withdrawal cognition) is somewhat broader than

turnover intention. The global-satisfaction subgroup of Tett and Meyer (1993) ($\rho = 0.72$, $SD_\rho = 0.13$, $k = 20$) leads to the same conclusions for satisfaction–commitment.

4.3 Results

Table 2 and Figure 2 report the comparisons (Steps 2 to 4). No correlation differs significantly from its benchmark mean, and every observed correlation lies inside its 95% prediction interval. On the evidence available, none of the six relationships among satisfaction, commitment and turnover intention is unusual for this group of expatriate educators.

Table 2: Comparison of Observed Correlations with Attenuated Benchmarks (n = 52)

	Observed			Against the Mean			Against the Distribution		
Relationship	**r**	**95% CI**	**ρ_{att}**	**z_M**	**p**	**p_{Holm}**	**95% PI**	**z_H**	**p**
Satisfaction–commitment	0.73	[0.57, 0.84]	0.61	+1.55	0.121	0.644	[0.38, 0.83]	+1.06	0.290
Satisfaction–intention	−0.26	[−0.50, 0.01]	−0.46	−1.61	0.107	0.644	[−0.74, −0.18]	−1.38	0.167
Commitment–intention	−0.46	[−0.65, −0.21]	−0.44	+0.18	0.856	1.000	[−0.76, −0.12]	+0.13	0.899
Affective–intention	−0.42	[−0.62, −0.17]	−0.48	−0.57	0.572	1.000	[−0.90, −0.07]	−0.30	0.761
Normative–intention	−0.35	[−0.57, −0.09]	−0.28	+0.54	0.589	1.000	[−0.77, 0.21]	+0.28	0.780
Continuance–intention	−0.33	[−0.55, −0.06]	−0.13	+1.49	0.135	0.644	[−0.47, 0.21]	+1.15	0.250

Note: $\rho_{att} = A \times \rho$, where $A = \sqrt{(r_{xx} \times r_{yy})}$ is the attenuation factor from the study's alpha coefficients (0.87, 0.86, 0.84, 0.86, 0.85 and 0.72 for the six rows in order). z_M tests the difference from the benchmark mean (Equation 4) and p_{Holm} adjusts its p value across the six tests. PI = 95% prediction interval for an observed correlation in a study of n = 52, allowing for between-study heterogeneity (Equation 6); z_H is the corresponding test (Equation 5). The sign of z is oriented so that positive values indicate an observed correlation stronger (further from zero) than the benchmark.

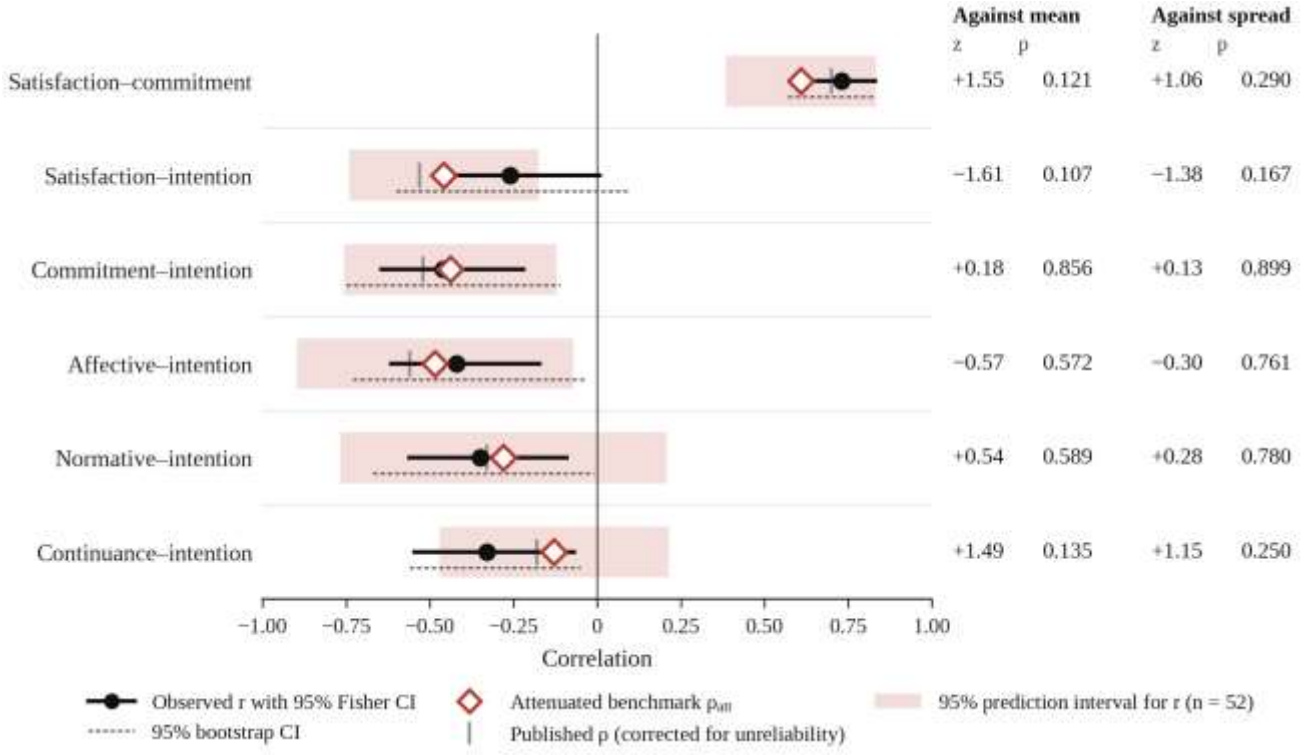
The individual comparisons are nonetheless instructive. **Attenuation can reverse an informal impression.** The satisfaction–commitment correlation of 0.73 looks "consistent with" the published 0.70. Once the benchmark is attenuated to the study's reliabilities (0.61), the observed correlation is in fact stronger than expected by about 0.12, although not significantly so ($p = 0.121$). This is exactly the kind of difference in the other direction that an informal comparison conceals.

Matching the construct can remove an apparent difference. The satisfaction–intention correlation (−0.26) is weaker than its attenuated, construct-matched benchmark (−0.46), but the difference is not significant ($p = 0.107$) and the benchmark lies inside the confidence interval. Section 4.5 shows how strongly this conclusion depends on analytic choices.

Close agreement. The overall commitment–intention correlation (−0.46) almost exactly matches its attenuated benchmark (−0.44), and the affective (−0.42 against −0.48) and normative (−0.35 against −0.28) component correlations are also close.

Continuance commitment. The continuance–intention correlation (−0.33) is stronger than its attenuated benchmark (−0.13). The difference is not significant (p = 0.135) and the observed value lies inside the prediction interval, but it is worth noting for two reasons. Continuance commitment usually has the weakest relationship with turnover cognitions of the three components [21], yet for expatriates the perceived costs of leaving, such as visas, relocation and children's schooling, may weigh more heavily than for domestic employees. At the same time, the continuance scale had the lowest reliability in the study (α = 0.61), which both attenuates its benchmark most and makes its estimate least precise. The observation is a hypothesis for future research rather than a finding.

Figure 2: Observed Correlations Compared with Attenuated Meta-Analytic Benchmarks (n = 52)



Note: Black circles and solid lines show each observed correlation and its 95% Fisher confidence interval; dashed lines show 95% percentile bootstrap intervals (2,000 resamples). Grey ticks mark the published benchmark ρ and open diamonds the benchmark attenuated to the study's reliabilities. Shaded bands are 95% prediction intervals for an observed correlation from a study of n = 52, allowing for between-study heterogeneity (Equation 6). The columns on the right give the tests against the benchmark mean (Equation 4) and against the distribution of true correlations (Equation 5); positive z indicates an observed correlation stronger than the benchmark.

4.4 Sensitivity of the Estimates

Table 3 reports the checks in Step 5, together with the detectability calculations from Step 6. The leave-one-out ranges and bootstrap intervals (2,000 resamples) are taken from the parent study [19].

Table 3: Sensitivity and Detectability of the Six Comparisons (n = 52)

		Sensitivity			Detectable Difference	
Relationship	**r**	**Leave-One-Out Range**	**95% Bootstrap CI**	**ρ_{att}**	**Weaker**	**Stronger**
Satisfaction–commitment	0.73	0.70 to 0.75	[0.57, 0.83]	0.61	0.31	0.19
Satisfaction–intention	−0.26	−0.44 to −0.20	[−0.60, 0.10]	−0.46	0.36	0.26
Commitment–intention	−0.46	−0.59 to −0.40	[−0.75, −0.11]	−0.44	0.37	0.26
Affective–intention	−0.42	−0.60 to −0.36	[−0.73, −0.03]	−0.48	0.36	0.25
Normative–intention	−0.35	−0.47 to −0.29	[−0.67, −0.01]	−0.28	0.39	0.32
Continuance–intention	−0.33	−0.38 to −0.25	[−0.56, −0.05]	−0.13	0.39	0.36

Note: Leave-one-out ranges and bootstrap intervals are from the parent study [19]; Fisher intervals are given in Table 2. In every row the attenuated benchmark lies inside the bootstrap interval. Detectable differences are the smallest true differences from ρ_{att}, in correlation units, that a study of n = 52 would detect with 80% power at $\alpha = 0.05$, two-sided (Equation 8).

Two points stand out. First, the satisfaction–commitment correlation is robust: removing any single respondent moves it by no more than 0.03, and its bootstrap and Fisher intervals agree closely. Second, every correlation involving turnover intention is sensitive to one respondent in particular, whose removal strengthens each of them. Without that respondent, the satisfaction–intention correlation would be −0.44, close to its attenuated benchmark of −0.46. The bootstrap intervals for these correlations are also noticeably wider than the Fisher intervals (for commitment–intention, [−0.75, −0.11] against [−0.65, −0.21]), indicating that the turnover-intention scores depart from bivariate normality. In such cases the Fisher interval is best treated as a lower bound on the uncertainty.

4.5 How Much Does the Conclusion Depend on Analytic Choices

The satisfaction–intention comparison shows how much the verdict can depend on choices made before any test is run. Table 4 starts from the analysis an informed researcher would most likely conduct, the pooled published benchmark attenuated to the study's reliabilities (row B), and changes one choice at a time.

Table 4: Satisfaction–Turnover Intention (r = −0.26): Conclusions under Alternative Analytic Choices

Analysis	ρ	ρ_{att}	Comparison	Result	Different
A. Informal comparison with the published pooled estimate	−0.58	—	Gap between point estimates	Gap of 0.32	Appears much weaker

B. Pooled estimate, attenuated (reference analysis)	−0.58	−0.50	z_M (Equation 4)	z = −2.00, p = 0.046	Yes, narrowly
C. As B, with Holm adjustment across six tests	−0.58	−0.50	p_{Holm}	p = 0.275	No
D. As B, using the percentile bootstrap interval	−0.58	−0.50	ρ_{att} inside [−0.60, 0.10]	Inside	No
E. As B, excluding one influential respondent (n = 51, r = −0.44)	−0.58	−0.50	z_M (Equation 4)	z = −0.55, p = 0.584	No
F. Construct-matched benchmark (turnover-intention scales)	−0.53	−0.46	z_M (Equation 4)	z = −1.61, p = 0.107	No
G. As F, allowing for between-study heterogeneity	−0.53	−0.46	z_H (Equation 5)	z = −1.38, p = 0.167	No
H. Mismatched benchmark (withdrawal-cognition scales)	−0.74	−0.64	z_M (Equation 4)	z = −3.43, p < 0.001	Yes, clearly

Note: Rows C to E each change one element of the reference analysis B; row G changes one element of row F. Negative z indicates an observed correlation weaker than the benchmark. Computed from the unrounded correlation, row B gives |z| = 2.01, p = 0.044. Row H is shown to illustrate the consequence of a construct mismatch; it is not an appropriate benchmark for a turnover-intention scale.

Under the reference analysis the shortfall is significant, but only narrowly. It does not survive adjustment for the six comparisons made (row C), a bootstrap interval that does not assume normality (row D), exclusion of a single respondent (row E), or replacement of the pooled benchmark with the subgroup that matches the study's measure (row F). Allowing for genuine between-study variation weakens it further (row G). Conversely, a researcher who unknowingly benchmarked against withdrawal-cognition scales would report a large and highly significant shortfall (row H), an artefact of construct mismatch rather than of the expatriate context. The defensible conclusion is that the evidence does not show satisfaction to be more weakly related to turnover intention among these educators than elsewhere. A modest weakening cannot be ruled out, but neither can the absence of any weakening.

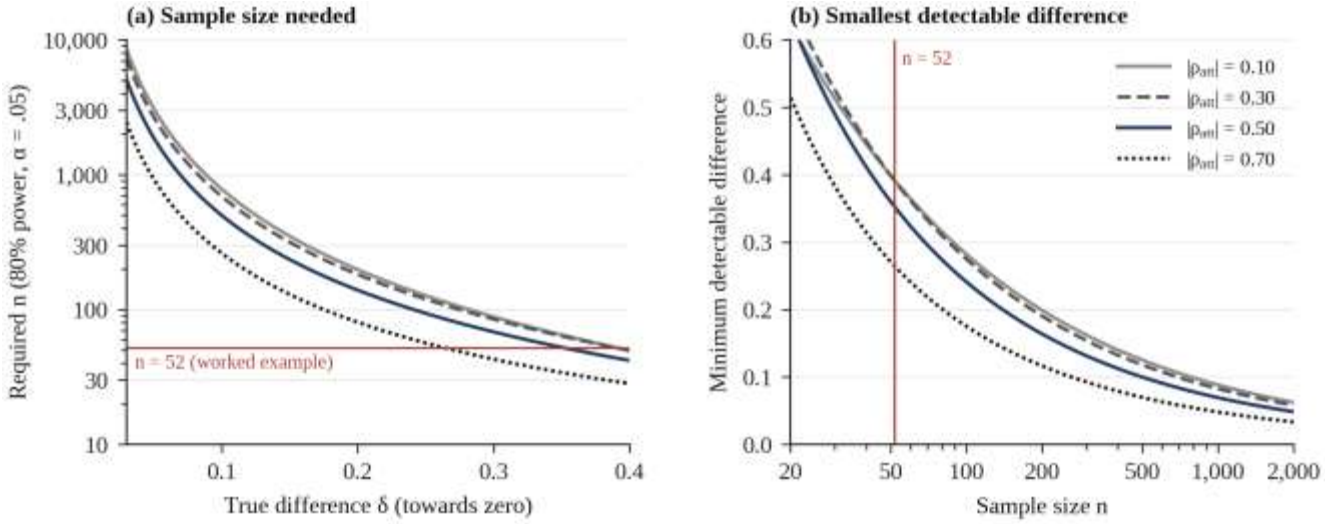
That conclusion leaves room for substantive interpretation without overstating it. The parent study found that satisfaction affected turnover intention mainly through affective commitment rather than directly [19]. Tett and Meyer (1993) [13], by contrast, found that satisfaction also contributes to turnover cognitions independently of commitment. If that direct path is weaker in this context, the bivariate satisfaction–intention correlation would be somewhat weaker than the benchmark, which is the direction of the point estimate. The benchmarking evidence is consistent with this interpretation but cannot confirm it; a larger sample would be needed to do so.

4.6 What the Study Could Have Detected

With n = 52, the smallest true departure from the satisfaction–intention benchmark (ρ_{att} = −0.46) detectable with 80% power is about 0.36 in the weaker direction and 0.26 in the stronger direction (Table 3). Across the six relationships, the corresponding figures range from 0.31 to 0.39 and from 0.19 to 0.36. Figure 3 generalises the point. Detecting a true weakening of 0.10 from a benchmark of

magnitude 0.50 would require about 500 respondents, a weakening of 0.20 about 140, and a weakening of 0.30 about 69 (Table 5).

Figure 3: Detectability of Departures from a Meta-Analytic Benchmark



Note: Panel (a) shows the sample size required to detect a true difference δ towards zero from an attenuated benchmark of magnitude $|\rho_{att}|$, with 80% power at α = 0.05, two-sided (Equation 7). Panel (b) shows the minimum detectable difference towards zero as a function of sample size (Equation 8). The red lines mark the worked example (n = 52).

A study of 52 can therefore detect only large departures from the literature. A non-significant difference means that any departure is unlikely to be very large, not that the relationship is the same as elsewhere. Claims of equivalence are harder still: with n = 52, the narrowest equivalence margin that the data could support, defined as the largest distance between ρ_{att} and the 90% confidence limits [18], ranges from 0.19 to 0.43 across the six relationships, far wider than any margin a researcher would wish to call "consistent with the literature".

5. Discussion

5.1 What the Example Shows

The worked example began with an informal comparison that suggested one striking discrepancy, a satisfaction–intention correlation less than half the size of the published estimate, and five relationships "consistent with the literature". The formal procedure changes both parts of that picture. Most of the striking discrepancy is explained by a mismatch of metric and construct, and what remains is too small and too fragile to distinguish from sampling error. "Consistent with the literature", meanwhile, turns out to mean something more modest than it sounds: the data are compatible with the benchmarks, but also with substantial departures from them in either direction.

Three general lessons follow. First, benchmark selection is an analytic decision whose consequences can exceed those of the statistical test. It should be justified and reported, and where several defensible benchmarks exist, the comparison should be shown under each. Second, "different from the average study" and "unusual among studies" are different questions. When meta-analyses report substantial heterogeneity, as is common in organisational research, a local result can differ from the mean for good

substantive reasons without being anomalous. Third, for a small study the most informative output of the procedure is often Step 6: a clear statement of the size of difference the study could and could not have detected.

5.2 Recommendations for Reporting

Researchers who benchmark small-sample correlations against meta-analytic estimates should do the following.

1. **Justify the benchmark.** State its source, the subgroup used, k, N, ρ and SD_ρ, and how closely its constructs and measures match the study's.
2. **Report intervals.** Give the 95% confidence interval for every observed correlation compared with a benchmark, and a bootstrap interval where normality is doubtful.
3. **Compare like with like.** State whether the benchmark is corrected for unreliability and, if so, attenuate it using the study's reliabilities.
4. **Test rather than eyeball, and keep the two questions apart.** Report the test against the benchmark mean and, where SD_ρ is available, the prediction interval; adjust for multiple comparisons.
5. **Report fragility.** State whether the conclusion changes when influential cases are removed or when a different defensible benchmark is used, and show both results.
6. **State what could have been detected.** Report the minimum detectable difference, so that non-significant results are not read as evidence of equivalence.

6. Limitations

The procedure has several limitations. The test against the benchmark mean treats ρ as known. With large meta-analyses this is reasonable, but for small subgroups (k = 9 for the commitment–intention benchmark, for example) the uncertainty in ρ itself is not negligible, and a fuller test would add the squared standard error of ρ to the variance in Equation 4. The heterogeneity-aware comparison in Equations 5 and 6 uses a normal approximation on the correlation scale; it is adequate for screening but approximate for correlations near ±1, and it inherits the imprecision of SD_ρ when k is small.

The attenuation formula assumes that the study's reliability estimates are accurate. With short scales and small samples they are themselves uncertain, as the three-item continuance scale (α = 0.61) illustrates, and coefficient alpha is a lower bound to reliability only under restrictive assumptions [23]. Meta-analyses may also apply different corrections (for range restriction or dichotomisation, for instance), so the methods of each source should be checked before attenuating. The two meta-analyses used here are also relatively old; more recent syntheses may offer benchmarks closer to current measures and populations.

Finally, the six comparisons in the worked example are not independent, because they share variables; Holm's procedure remains valid under dependence but is conservative. The leave-one-out and bootstrap results come from the parent study, and the procedure cannot by itself say whether an influential respondent reflects an error, a legitimate extreme case or a member of a different population. That judgement requires knowledge of the data.

7. Conclusion

Comparing small-sample results with meta-analytic benchmarks is natural and valuable, but done informally it tends to exaggerate differences and to overstate agreement. The six-step procedure set out here (select a commensurable benchmark, estimate the interval, align the metrics, compare with both the

mean and the distribution, stress-test, and calibrate) gives researchers a defensible way to say whether a result is unusual. In the worked example none of the six relationships was, and the one that appeared extraordinary owed its apparent distinctiveness more to the choice of benchmark than to the educators surveyed. That is a less dramatic story than an informal comparison would tell, and a more honest one.

8. Appendix

8.1 Sample Sizes for Detecting Departures from a Benchmark

Table 5 gives the number of respondents needed to detect a true difference δ towards zero from an attenuated benchmark, computed from Equation 7 and rounded up. Detecting a difference of the same size away from zero requires somewhat fewer respondents, because the Fisher transformation stretches the correlation scale near ±1.

Table 5: Respondents Needed to Detect a True Difference towards Zero (80% Power, α = 0.05, Two-Sided)

$\lvert\rho_{att}\rvert$	δ = 0.05	δ = 0.10	δ = 0.15	δ = 0.20	δ = 0.30
0.10	3,107	783	351	198	89
0.20	2,952	752	340	194	89
0.30	2,685	692	316	183	85
0.40	2,320	606	281	164	79
0.50	1,884	501	236	140	69
0.60	1,408	383	184	112	57
0.70	931	262	130	81	43

8.2 Spreadsheet Implementation

The formulas in Table 6 implement every step in Microsoft Excel or any compatible spreadsheet. The inputs are entered in cells B2 to B7: the observed correlation r (B2), the sample size n (B3), the published benchmark ρ (B4), its SD_ρ (B5), and the study's reliabilities r_{xx} (B6) and r_{yy} (B7). Each formula is placed in the cell shown so that later formulas can refer to it.

Table 6: Spreadsheet Formulas for the Six-Step Procedure

Quantity	Equation	Cell	Formula
Fisher z of r	1	B8	=FISHER(B2)
Lower 95% confidence limit	2	B9	=FISHERINV(B8-1.96/SQRT(B3-3))
Upper 95% confidence limit	2	B10	=FISHERINV(B8+1.96/SQRT(B3-3))
Attenuated benchmark ρ_{att}	3	B11	=B4*SQRT(B6*B7)
Attenuated SD_{att}	3	B12	=B5*SQRT(B6*B7)
Test against the mean, z_M	4	B13	=(B8-FISHER(B11))*SQRT(B3-3)

Two-sided p for z_M	4	B14	=2*(1-NORM.S.DIST(ABS(B13),TRUE))
Denominator of Equation 5	5	B15	=SQRT(B12^2+(1-B11^2)^2/(B3-1))
Test against the distribution, z_H	5	B16	=(B2-B11)/B15
Two-sided p for z_H	5	B17	=2*(1-NORM.S.DIST(ABS(B16),TRUE))
Lower 95% prediction limit	6	B18	=B11-1.96*B15
Upper 95% prediction limit	6	B19	=B11+1.96*B15
Alternative value ρ_{att} + δ (input)	7	B20	=B11+0.10
Required n (80% power)	7	B21	=ROUNDUP(((NORM.S.INV(0.975)+NORM.S.INV(0.8))/ABS(FISHER(B20)-FISHER(B11)))^2+3,0)
Lower detectable limit at n	8	B22	=FISHERINV(FISHER(B11)-2.8/SQRT(B3-3))
Upper detectable limit at n	8	B23	=FISHERINV(FISHER(B11)+2.8/SQRT(B3-3))

Note: For the Holm adjustment, the k two-sided p values are sorted from smallest to largest, the i-th smallest is multiplied by (k − i + 1), and each value is then replaced by the largest value at or before it in the sorted sequence, capped at 1.

9. Data Availability and Ethics

The worked example uses summary statistics from the author's doctoral thesis [19]. The benchmark comparisons, heterogeneity analyses and power calculations are new to this paper and can be reproduced from the values in Tables 1 to 3 using the formulas in Table 6.

The study was approved through the research ethics procedures of Westford University College and UCAM on 1 July 2026. Participation was voluntary and anonymous, and all participants gave informed consent before taking part.

10. References

1. F.A. Bosco, H. Aguinis, K. Singh, J.G. Field, C.A. Pierce, "Correlational Effect Size Benchmarks", Journal of Applied Psychology, 2015, 100 (2), 431–449.
2. J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd Edition, Lawrence Erlbaum Associates, Hillsdale, 1988.
3. G.E. Gignac, E.T. Szodorai, "Effect Size Guidelines for Individual Differences Researchers", Personality and Individual Differences, 2016, 102, 74–78.
4. F.D. Schönbrodt, M. Perugini, "At What Sample Size Do Correlations Stabilize?", Journal of Research in Personality, 2013, 47 (5), 609–612.
5. G. Cumming, "The New Statistics: Why and How", Psychological Science, 2014, 25 (1), 7–29.
6. R.A. Fisher, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population", Biometrika, 1915, 10 (4), 507–521.
7. R.A. Fisher, "On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample", Metron, 1921, 1, 3–32.

8. C. Spearman, "The Proof and Measurement of Association between Two Things", American Journal of Psychology, 1904, 15 (1), 72–101.
9. F.L. Schmidt, J.E. Hunter, Methods of Meta-Analysis: Correcting Error and Bias in Research Findings, 3rd Edition, Sage Publications, Thousand Oaks, 2015.
10. E.M. Whitener, "Confusion of Confidence Intervals and Credibility Intervals in Meta-Analysis", Journal of Applied Psychology, 1990, 75 (3), 315–321.
11. J.P.T. Higgins, S.G. Thompson, D.J. Spiegelhalter, "A Re-Evaluation of Random-Effects Meta-Analysis", Journal of the Royal Statistical Society: Series A (Statistics in Society), 2009, 172 (1), 137–159.
12. J. IntHout, J.P.A. Ioannidis, M.M. Rovers, J.J. Goeman, "Plea for Routinely Presenting Prediction Intervals in Meta-Analysis", BMJ Open, 2016, 6 (7), e010247.
13. R.P. Tett, J.P. Meyer, "Job Satisfaction, Organizational Commitment, Turnover Intention, and Turnover: Path Analyses Based on Meta-Analytic Findings", Personnel Psychology, 1993, 46 (2), 259–293.
14. S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure", Scandinavian Journal of Statistics, 1979, 6 (2), 65–70.
15. B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, 1993.
16. A.J. Bishara, J.B. Hittner, "Confidence Intervals for Correlations When Data Are Not Normal", Behavior Research Methods, 2017, 49 (1), 294–309.
17. L.J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests", Psychometrika, 1951, 16 (3), 297–334.
18. D. Lakens, "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses", Social Psychological and Personality Science, 2017, 8 (4), 355–362.
19. I. Quarmson, Addressing Expatriate Educator Turnover in Guangzhou International High Schools: Investigating the Influence of Job Satisfaction and Organisational Commitment toward the Development of a Strategic Employee Retention Framework, Doctor of Business Administration Thesis, Westford University College and Universidad Católica San Antonio de Murcia, 2026. (Unpublished)
20. J.P. Meyer, N.J. Allen, "A Three-Component Conceptualization of Organizational Commitment", Human Resource Management Review, 1991, 1 (1), 61–89.
21. J.P. Meyer, D.J. Stanley, L. Herscovitch, L. Topolnytsky, "Affective, Continuance, and Normative Commitment to the Organization: A Meta-Analysis of Antecedents, Correlates, and Consequences", Journal of Vocational Behavior, 2002, 61 (1), 20–52.
22. N.J. Allen, J.P. Meyer, "The Measurement and Antecedents of Affective, Continuance and Normative Commitment to the Organization", Journal of Occupational Psychology, 1990, 63 (1), 1–18.
23. K. Sijtsma, "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha", Psychometrika, 2009, 74 (1), 107–120.

Licensed under Creative Commons Attribution-ShareAlike 4.0 International License