

# Pre-Submission Methods Review

## Sample report

### About this sample

This is a complete report, not an excerpt. It is the document a buyer receives, with nothing removed and nothing softened.

The manuscript reviewed here is the reviewer's own unpublished working paper. That is deliberate. A confidential client manuscript is never used in any sample, page or talk, which is a commitment this service makes in writing, so the only manuscript available to demonstrate the product is one the seller owns outright.

**A reviewer cannot referee his own paper, and this sample does not pretend otherwise.** It is written as the report an external methods referee would write. The objections below are real ones; several of them will change the paper, and the two largest have not yet been answered.

**The target venue was chosen for this demonstration.** The paper is a working paper and is not under submission anywhere. Journal of Business Ethics was named as a realistic destination for a manuscript of this kind so that the desk-reject screen has something to screen against. Every venue requirement in Section 2 was read from that journal's published submission guidelines on 24 September 2026 and is cited as read on that date.

## 0. Title block

|                      |                                                                                                                                                                                                                                                                  |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Manuscript           | How Many Firms Can an ESG Rating Tell Apart? ESG Rating Divergence as Reliability, and the Question Regulation (EU) 2024/3005 Does Not Ask                                                                                                                       |
| Author               | Milos Kankaras                                                                                                                                                                                                                                                   |
| Target venue         | Journal of Business Ethics (Springer), Original paper                                                                                                                                                                                                            |
| Design as classified | Secondary analysis of a published correlation matrix, with simulation. Not an empirical study of new data; not experimental. Causal-language flags are therefore not applied, and inferential claims are read as claims about what the published matrix implies. |
| Length               | 5,841 words in total; 5,471 excluding the title page, abstract and keywords; 19 references; 3 tables                                                                                                                                                             |
| Reviewer             | Dr. Milos Kankaras, ORCID 0000-0002-3190-7751                                                                                                                                                                                                                    |
| Date                 | 24 September 2026                                                                                                                                                                                                                                                |
| Tier                 | Tier 1, report only                                                                                                                                                                                                                                              |

*If the design classification above is wrong, stop reading and say so. It governs which of the objections below are defects.*

## 1. Summary judgement

The paper's arithmetic is correct and I could not break it. I recomputed the Spearman-Brown figure, the standard error, the minimum significant difference and the rank-order correlations from the published matrix, and each reproduces to the digits reported. The problems are not in the calculations. They are in what the calculations are called, in what is not reported alongside them, and in the fit between this manuscript and the journal named on the submission.

**Recommendation: desk-reject risk at this venue, and major revision at any venue.**

Two things drive that. First, the manuscript will very probably not reach a reviewer at Journal of Business Ethics. It engages no business-ethics literature, it is some 2,500 words shorter than that journal's stated range for an Original paper, and it carries the author's name, affiliation, ORCID, email and personal website into a double-anonymous review. Any one of those three is a plausible return; together they are close to decisive. Section 2 lists what would have to change, and Section 6 says where I would send it instead.

Second, and this survives any change of venue, the paper computes a quantity from disagreement between providers and then reports it in the vocabulary of error: "standard error of measurement", "true standing", "hit rate". Section 2 of the manuscript concedes, correctly and in its own words, that most of the residual variance is not error but provider-specific construct, and cites the source that demonstrates it. Sections 3 and 5 then proceed as though the concession had not been made. A referee who reads Section 2 carefully will use it against Section 3, and the concession will not protect the paper because it is the paper's own.

The third issue is smaller but will sting. A manuscript whose central argument is that providers should publish a reliability coefficient and a standard error reports its own central reliability estimate without an interval of any kind.

## 2. Desk-reject screen: Journal of Business Ethics

Every requirement below was read from the journal's published submission guidelines on 24 September 2026. Rows marked Unverified are ones I could not confirm from that source and are flagged rather than guessed.

| Check             | What the venue requires                                                                                                                                                   | What the manuscript does                                                                                                                                                   | Verdict               |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Scope fit         | A journal of business ethics, organised in more than thirty sections; authors advised to identify their section before submitting                                         | No engagement with any business-ethics literature or theory. The 19 references are finance, psychometrics, EU law and classical test theory. No section is an evident home | Blocking              |
| Article type      | Author must declare one of: Original paper, Review essay, Commentary                                                                                                      | Not declared. At 5,471 words with a "the question the Regulation does not ask" framing, an editor may read it as a Commentary, which is a different bar                    | Fix before submitting |
| Contribution type | "We do not consider papers whose only contributions are bibliometric or a response to another paper"; a response developed into an independent contribution is considered | The analysis rests entirely on one published table from Berg et al. (2022). The independent contribution is real but is nowhere argued as such                             | Fix before submitting |
| Length            | Approximately 8,000 to 10,000 words including references; no longer than                                                                                                  | 5,841 words in total, 5,471 excluding the title page and abstract. Roughly 2,500 words below the                                                                           | Fix before submitting |

| Check                           | What the venue requires                                                                                                                                                                               | What the manuscript does                                                                                                                                                                                                                                                                                                                                                 | Verdict                                    |
|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|
|                                 | 12,000                                                                                                                                                                                                | bottom of the stated range                                                                                                                                                                                                                                                                                                                                               |                                            |
| Brief notes                     | "The general rule is that we do not publish brief discussion notes"                                                                                                                                   | The combination of the length above and the commentary framing invites exactly this reading                                                                                                                                                                                                                                                                              | Fix before submitting                      |
| Abstract                        | 150 to 250 words                                                                                                                                                                                      | 240 words. Within range                                                                                                                                                                                                                                                                                                                                                  | Clear                                      |
| Keywords                        | 4 to 6                                                                                                                                                                                                | 7                                                                                                                                                                                                                                                                                                                                                                        | Fix before submitting                      |
| Double-anonymous review         | Author must anonymise the manuscript and all accompanying files; names, affiliations and other identifying information removed to a separate title page; self-citation in the third person            | Byline carries the author's name and affiliation; the header carries the ORCID, a correspondence address at a named organisation, and a personal website. Section 1 reads "our companion paper on international rankings examines (Kankaras, 2026)", the first-person self-citation the guidelines specifically prohibit, and the companion paper is named in the header | Blocking                                   |
| Pre-publication posting         | Authors should, where possible, remove or refrain from posting pre-publication versions on publicly accessible sites; where not possible, use a significantly different title to reduce searchability | The paper's own dissemination plan is to post it as an SSRN working paper under this title. That plan and this submission are in direct conflict and one of them has to give                                                                                                                                                                                             | Blocking, and a decision rather than a fix |
| Statements and Declarations     | A section headed "Statements and Declarations" including Competing Interests; "submissions that do not include relevant declarations will be returned as incomplete"                                  | A section headed "Declarations" states no competing interests and no funding. The content is right and the heading is not                                                                                                                                                                                                                                                | Fix before submitting                      |
| Large language model use        | LLM use should be documented in the Methods section, or a suitable alternative part of the manuscript, where it goes beyond AI-assisted copy editing                                                  | The disclosure names drafting assistance and analysis-script checking, which is beyond copy editing, and it sits on the title page. The title page is not sent to reviewers, so under this journal's own process no reviewer sees it                                                                                                                                     | Fix before submitting                      |
| Reference style                 | APA 7, name and year in text, DOIs as full links                                                                                                                                                      | Followed throughout. DOIs present on all but the two arXiv and EUR-Lex items, which carry stable URLs                                                                                                                                                                                                                                                                    | Clear                                      |
| Tables                          | Arabic numerals, cited in text in consecutive order, each captioned                                                                                                                                   | Three tables, numbered, cited in order, each captioned with a substantive note                                                                                                                                                                                                                                                                                           | Clear                                      |
| Headings                        | No more than three levels                                                                                                                                                                             | Two levels used                                                                                                                                                                                                                                                                                                                                                          | Clear                                      |
| File format                     | Word or LaTeX, editable source at every submission                                                                                                                                                    | DOCX built                                                                                                                                                                                                                                                                                                                                                               | Clear                                      |
| Ethics approval and consent     | Required where applicable                                                                                                                                                                             | No human participants and no new data; correctly absent                                                                                                                                                                                                                                                                                                                  | Clear                                      |
| Data availability               | Journal research-data policy applies                                                                                                                                                                  | A data and code availability section names a Zenodo deposit with the DOI still to be inserted. Submitting with an uninserted DOI is a returnable incompleteness                                                                                                                                                                                                          | Fix before submitting                      |
| Section selection at submission | Author advised to identify one of the thirty-plus sections                                                                                                                                            | I could not determine from the guidelines whether the section is a required submission field or an advisory step, and the section list is on a separate page I did not read                                                                                                                                                                                              | Unverified                                 |
| Response to a published paper   | A response developed into an independent contribution is considered                                                                                                                                   | Whether a section editor would classify this as a response to Berg et al. at all is an editorial judgement. I am reporting the risk, not the outcome                                                                                                                                                                                                                     | Unverified                                 |

**Three blocking rows and eight fixes.** The anonymisation and the declarations heading are half an hour of work. The length, the scope fit and the SSRN conflict are not; they are decisions about what this paper is and where it goes, and they are taken up in Section 6.

### 3. Objections, severity-ranked

Ordered by the probability that a referee blocks on them, not by the order they appear in the manuscript.

#### [1] Major - The residual is called error, and the paper has already conceded it is not

**Locus.** Section 2, the paragraph beginning "What the coefficient measures needs stating precisely": "The remaining 46 per cent mixes provider-specific construct, which Berg et al.'s decomposition shows to be partly systematic, with error, and this matrix cannot separate the two." Then Section 3, second paragraph: "The standard error of measurement is the standard deviation of the error in one rating".

**What the referee will say.** *"The author concedes in Section 2 that the complement of the reliability coefficient confounds systematic provider-specific construct variance with random error, and that this matrix cannot separate them. Section 3 then treats the whole of that complement as error: it takes its square root, calls the result a standard error of measurement, and builds every headline quantity in the paper on it, the 9th-to-91st percentile band, the top-decile hit rate, the number of distinguishable bands. Berg, Koelbel and Rigobon's own decomposition, which the author cites approvingly, attributes 56 per cent of divergence to differences in what is measured and 38 per cent to differences in coverage. Those are not error. A standard error of measurement computed from systematic construct differences is not a standard error of measurement. The author cannot both make the concession and keep the arithmetic."*

**Why it is raised.** This is the load-bearing step of the paper and it is the one place where the manuscript's own text supplies the refutation. The concession in Section 2 is precise and correct, and the language of Section 3 is inconsistent with it. The manuscript does defend itself, in the same paragraph, by reframing the quantity as "the cross-provider case, the spread of placements a user meets across providers of equal typicality". That defence is sound and it is the right one. But it is made once, in a subordinate clause, four pages before the quantities it licenses, and it is not carried into the vocabulary. "True standing", in Table 2's stub and note, is the specific phrase that cannot survive: under the reframing there is no true standing, there is a consensus placement.

**Verdict.** Real defect. Not in the arithmetic, which is right under a clearly stated interpretation, but in the reporting, which uses a different interpretation's vocabulary throughout. A referee will read this as the author wanting the rhetorical force of measurement error while having disclaimed it.

**Fix.** Do not weaken the result. Rename it, once, and then be consistent.

- Introduce the quantity in Section 3 under a name that says what it is. "Cross-provider dispersion", or "interchangeability error", with one sentence defining it as the standard deviation of the placements a firm of given consensus standing receives across providers, and

one sentence saying that it is an upper bound on the provider's own re-rating error rather than an estimate of it.

- Replace "true standing" everywhere with "consensus standing": Table 2's stub, Table 2's note, Table 3's hit-rate definition, and the three occurrences in Section 3's prose.
- Keep "standard error of measurement" only in Section 4, where it names the quantity the Standards require a provider to report and is the correct term. The distinction between the two is one of the paper's better points and at present the shared vocabulary hides it.
- Add one sentence to the abstract making the reframing visible there.

**If unfixed.** A methods-literate referee rejects, or demands the change as a condition. This is the objection I would lead with if I were reviewing it for a journal, and I would not accept a revision that answered it only in a response letter.

## [2] Major - The resolution index is invented, the established index gives half the answer, and the abstract reports the larger number

**Locus.** Section 3, the paragraph beginning "The question a user can meaningfully ask is how many bands a rating can distinguish", and Table 3, column "Resolution, 1,000 firms". Carried into the abstract as "one rating would resolve a thousand-firm universe into three or four bands" and into the conclusion unqualified.

**What the referee will say.** *"The author defines a new index, the expected range of the universe divided by the minimum significant pairwise difference, and reports its value as the paper's headline finding. No precedent is cited and none is compared. But this quantity has a standard form in the measurement literature the author is otherwise drawing on: the separation index and the associated number of distinct strata,  $G$  equals the root of reliability over one minus reliability, with strata equal to four  $G$  plus one, over three. At the paper's own coefficient of 0.541 that gives  $G$  of 1.09 and 1.8 strata. The author's index gives 3.5. The paper reports the figure that is roughly twice the established one, without mentioning that the established one exists."*

**Why it is raised.** I checked this. At the manuscript's own reliability of 0.541 the separation index is 1.086 and the strata figure is 1.78. At the six-rater average of 0.88 the strata figure is 3.94 against the manuscript's 6.6. The gap is systematic and large, and it runs in the direction that makes the paper's claim weaker in every cell. The manuscript is not hiding it: it reports in the same paragraph that under a conservative non-overlap criterion the thousand-firm figure is 2.4 rather than 3.5, and it says its own index "carries no multiple-comparison control and is an upper bound". That conservative 2.4 is much closer to the strata figure of 1.8, which suggests the two approaches broadly agree and the liberal criterion is the outlier. But it is the liberal figure that reaches the abstract, the title and the conclusion, and the qualification does not travel with it. There is a second problem inside the index: counting mutually distinguishable firms by pairwise tests at the 5 per cent level across a thousand firms is roughly half a million comparisons.

**Verdict.** Real defect, and the most fixable of the three majors.

**Fix.** Compute and report the separation index and the strata figure alongside the paper's own index, as two extra columns in Table 3, citing the measurement source rather than deriving it afresh.

- Where they disagree, say why. The honest account is that the paper's index measures the observed range against a pairwise significance criterion while the strata index measures the

spread of estimated true standings against the error, and the second is the more conservative and the more conventional.

- Move the conservative non-overlap figure out of the qualifying sentence and into the headline. State the range: a single rating resolves a large universe into roughly two to four bands depending on the criterion. That is still a striking claim, it is robust across all three criteria, and it cannot be attacked the way "three or four" can.
- Change the abstract and the conclusion accordingly. At present both state the most generous figure the paper produced.

**If unfixed.** A psychometric referee finds the strata index in ten minutes, as I did, and the review says the author reported the flattering number. That is a much worse outcome than reporting a range, and the paper does not need the larger figure to make its point.

### [3] Major - A paper demanding that providers report uncertainty reports none of its own

**Locus.** Everywhere. Section 2: "That estimate is 0.54." The six one-factor communalities, 0.25 to 0.74. Table 3, all ten rows. Section 3's simulation, "2,000 universes per cell". Not one interval appears in the manuscript.

**What the referee will say.** *"The paper's recommendation is that providers publish a reliability coefficient and a standard error. Its own reliability estimate is reported as a point value, as are the six provider-level communalities, the standard error of measurement, the rank standard deviations, the hit rates and the resolution figures. The simulation reports no Monte Carlo error. A reviewer is entitled to ask the author to meet the standard the paper is proposing."*

**Why it is raised.** It is correct, it is easy to fix, and the irony will be irresistible to a referee. It is also substantively important for the provider-level results: the communalities span 0.25 to 0.74 and the paper draws a contrast between MSCI and Moody's ESG from them. Loadings estimated from a six-variable matrix are not precise, and without intervals the reader cannot tell whether that contrast is real or whether the six communalities are statistically indistinguishable. My own reading is that the MSCI-versus-Moody's contrast will survive and several of the middle comparisons will not, but the manuscript gives no way to check. One complication should not be papered over: the matrix is reproduced from a published table rounded to two decimals, so intervals computed from it carry the rounding as well as the sampling error. That is a reason to state the caveat, not to omit the intervals.

**Verdict.** Real defect.

**Fix.** Bootstrap the whole pipeline from the published matrix at  $N = 924$  and report a 95 per cent interval for the mean correlation, the six communalities, the six-rater omega, and the resolution figure at the mean coefficient. The analysis is already scripted and deposited, so this is a day of work at most.

- Report the Monte Carlo standard error on the simulated rank standard deviations and hit rates, or raise the replication count until it is negligible and say so.
- Add one sentence recording that the intervals are conditional on a matrix published to two decimals.
- Then say so in the conclusion. A paper that reports its own uncertainty while asking providers to report theirs is making its argument twice.

**If unfixed.** Minimally, an embarrassing reviewer comment. More likely, a demand for exactly this analysis as a condition of revision, which costs a review round.

#### [4] Major - The heterogeneity test treats fifteen dependent correlations as independent

**Locus.** Section 2: "The data reject strict parallelism, since the fifteen correlations spread far more than sampling error allows (standard deviation 0.12 against a sampling standard error of about 0.02 each)".

**What the referee will say.** *"The fifteen pairwise correlations are computed on one sample of 924 firms across six raters. They are not independent: each rater appears in five of them, and the correlations share firms throughout. Comparing the observed standard deviation of fifteen dependent statistics against the sampling standard error of a single independent correlation is not a test of anything. The conclusion may well be right, but this is not the evidence for it."*

**Why it is raised.** It is a genuine statistical error rather than a matter of emphasis, and it sits in the sentence that licenses the move away from the parallel model. The comparison as written has no null distribution behind it. The conclusion almost certainly survives, and the manuscript already holds the better evidence for it: the one-factor communalities run from 0.25 to 0.74, which is incompatible with parallelism on its face, and the chi-square rejects. So this is a defect in the argument's plumbing rather than in its result.

**Verdict.** Real defect, and a cheap one.

**Fix.** Either drop the comparison and rest the claim on the one-factor solution, which the paper reports two paragraphs later and which is the appropriate evidence, or run a proper test of equality of dependent correlations and report it. The first is the better move: the sentence is doing rhetorical work the model already does substantively.

**If unfixed.** A statistical referee names it, and it is the kind of error that changes how the rest of the paper is read even though nothing else depends on it.

#### [5] Minor, close to major - The factor model is fitted to a rounded published matrix, and reported without fit indices

**Locus.** Section 2: "Fitted by maximum likelihood, the model misfits (chi-square 131.8 on 9 df)... Freeing that residual covariance (0.24) brings the root mean square residual to 0.02 (chi-square 25.7 on 8 df, which still rejects at this sample size but leaves no residual above 0.05)".

**What the referee will say.** *"Three things. The model is fitted to a correlation matrix reproduced from a published table at two decimal places and treated as though it were the sample matrix from 924 firms; at this sample size the rounding is not negligible for a chi-square. No fit indices are reported beyond the chi-square and a residual summary: CFI, TLI, RMSEA with its interval, and SRMR are the conventional set and their absence is conspicuous in a paper about measurement standards. And freeing a residual covariance after seeing the misfit is a post-hoc respecification; the substantive justification offered is persuasive, but the respecification is exploratory and the manuscript does not say so."*

**Why it is raised.** All three are correct. The third is the most consequential, because the freed KLD-MSCI residual is what produces the reported communalities, and those carry the provider-level argument and the 0.25-to-0.74 range quoted in the abstract. The substantive story behind the freed parameter is good: KLD had been an MSCI database since 2010, and the manuscript cites the category-level correlation of 0.69 for the pair as corroboration. That is a genuine prior justification rather than a fishing expedition, and it should be presented as one.

**Verdict.** Real defect in reporting, defensible in substance.

**Fix.** Report the full fit index set for both solutions.

- State that the matrix is published to two decimals and give a sensitivity check, which can be as simple as perturbing each correlation within its rounding interval and re-fitting.
- Say explicitly that the shared-house residual was specified on the prior ground of the KLD-MSCI ownership relation and corroborated by the category-level figure, rather than discovered in the residuals, if that is the true order of events. If it is not, say that instead and label the solution exploratory.

**If unfixed.** A statistical referee asks for the fit indices as a matter of course, which is a revision round spent on something that could have been in the first submission.

## [6] Minor - Normal-theory bands rest on an input the paper itself describes as an integer count

**Locus.** Table 2's note: "Interval is the normal-distribution function of the root of the reliability times the observed standard score". Table 1's note: "All six inputs are numeric scores, KLD's an integer count of strengths minus weaknesses." Section 3: "These figures are exact under normality where closed forms exist".

**What the referee will say.** *"Every percentile band in Table 2 and every resolution figure in Table 3 assumes normally distributed true standings and errors with common variance. The paper says so and says it has not examined alternatives. But its own Table 1 note records that one of the six inputs is an integer count of strengths minus weaknesses, which is bounded, discrete and in published samples right-skewed, and the Pearson correlations involving it are attenuated by that metric before any of this arithmetic begins. The assumption is not a formality here; it is contradicted by a feature the manuscript has already documented."*

**Why it is raised.** The manuscript names the assumption honestly, which is why this is minor rather than major. But naming an assumption is not the same as testing it when the paper has the material to test it, and a simulation framework is already built and deposited.

**Verdict.** Real defect, small and cheap to close.

**Fix.** Re-run the existing simulation under two or three non-normal true-score distributions, skewed and heavy-tailed, and report whether the resolution figures and hit rates move. My expectation is that they move little and the paper gains a robustness paragraph for a few hours of compute. If they move a lot, the paper needed to know.

- Add one sentence acknowledging the KLD metric and its attenuating effect on the five correlations that involve it.

**If unfixed.** Survivable. A careful referee asks for the robustness check and the paper supplies it in revision.

### [7] Minor - The 2014 vintage is caveated four times and answered once

**Locus.** Section 1: "The correlations are for 2014, and the providers have since merged, been renamed and rebuilt their methods". Section 3's closing paragraph. Section 5: "The evidence is a 2014 matrix of 924 large firms".

**What the referee will say.** *"The paper draws policy conclusions about a Regulation applying from July 2026 from a correlation matrix of 2014, in a market that has consolidated and re-tooled in the interval. The 2017 replication is nine years old. The author acknowledges this repeatedly and then recommends that somebody else do the current computation."*

**Why it is raised.** The paper's claim is explicitly about what the published evidence implies, the caveat is stated in the introduction and repeated wherever it bites, and the current data are proprietary and expensive. That is a complete answer and it should be made once, prominently, rather than four times apologetically. A referee who sees the limitation owned once and answered is satisfied; a referee who sees it hedged four times concludes the author is uneasy.

**Verdict.** Defensible. Drafted response in Section 5.

**Fix.** Consolidate the four caveats into one signposted statement early in the paper.

- Strengthen the Section 5 recommendation from "any academic team could repeat" into a specific, costed proposal naming the data source and the party best placed to bear it.

**If unfixed.** Nothing, if the response letter makes the argument. Presentational only.

### [8] Minor - The within-industry coefficient is the one users need and the paper does not have it

**Locus.** Section 3's closing paragraph: "within-industry agreement, on which best-in-class selection depends, is not published"; and Section 3's opening: "Best-in-class selection ranks firms within an industry peer group of a few dozen".

**What the referee will say.** *"The paper identifies best-in-class selection within industry peer groups as a principal use of ESG ratings, shows that a group of fifty resolves into two or three bands, and then concedes that the within-industry coefficient it would need is unpublished, substituting a random draw of fifty from the pooled universe. The substitution is not innocuous: within an industry, firms are more alike on the attributes providers score, so the range is compressed, and providers' industry adjustments are shared, so errors may be positively correlated. Both push in directions the paper has not signed."*

**Why it is raised.** Section 3 does set out the direction of the bias for the absolute-versus-relative cases and Table 3's note flags the group of fifty as an upper bound. The weakness is placement: the argument is in a note and a closing sentence, and the "two or three bands for a group of fifty" figure travels without it.

**Verdict.** Defensible, and the manuscript already argues it well. Drafted response in Section 5.

**Fix.** Promote the bias reasoning into the text at the point the figure is first given, and relabel the Table 3 column to make the random-draw basis explicit.

**If unfixed.** A referee reads the figure without its qualification and treats it as an over-claim.

### [9] Minor - The abstract and the tables headline different numbers for the same quantity

**Locus.** Abstract: "Averaging the six raises reliability to 0.86." Table 2, fourth column header: "reliability 0.88 (six-rater average, parallel)". Table 3, final row: 0.88.

**What the referee will say.** *"The abstract reports 0.86 for the six-rater average and both tables headline 0.88. Which is it?"*

**Why it is raised.** The manuscript is internally consistent and both table notes explain that 0.88 is the parallel upper bound and 0.86 the figure with the KLD-MSCI residual freed. But a reader meets 0.86 in the abstract and 0.88 in both tables, and a referee skimming for consistency between text and tables will flag it before reading the notes.

**Verdict.** Real, and purely cosmetic.

**Fix.** Use 0.86 as the headline figure everywhere, since it is the one the paper's preferred model produces, and show 0.88 as the parallel bound in the same cell. One edit per table plus the column header.

**If unfixed.** A one-line reviewer comment and a small loss of confidence.

### [10] Note - The Zenodo DOI is unresolved in a submission-ready file

**Locus.** Title page: "Data and code: Zenodo, DOI to be inserted at deposit." Data and code availability: "deposited on Zenodo under the DOI given on the title page".

**What the referee will say.** *"The data availability statement points at a DOI that does not exist yet."*

**Why it is raised.** It interacts with the anonymisation problem: a Zenodo deposit under the author's name, cited in a double-anonymous manuscript, deanonymises the paper. The deposit has to appear on the title page only, which is the arrangement the companion paper used at its own submission.

**Verdict.** Real, and blocking at submission rather than at review.

**Fix.** Deposit, insert the DOI on the title page only, and keep it out of the anonymised manuscript body.

**If unfixed.** Returned as incomplete by the editorial office.

### [11] Note - Article 15(7) is quoted but never turned on the paper's own recommendation

**Locus.** Section 4: "Article 15(7) requires methodologies 'rigorous, systematic, independent and capable of justification'".

**What the referee will say.** *"If Article 15(7) already requires methodologies capable of justification, why does the paper rest its ask on the limitations item at Annex III point 1(m) rather than on 15(7)?"*

**Why it is raised.** Observation rather than objection. The paper makes a strong case that 15(7)'s language could carry a reliability requirement without amendment, and then does not pursue it. The 15(7) route is the more interesting argument and a legal-minded reader will notice it is dropped.

**Verdict.** Not a defect. One paragraph, and it strengthens Section 4's conclusion.

**Fix.** Add the paragraph.

**If unfixed.** Nothing. An opportunity rather than a risk.

## [12] Reviewer-noise - "Psychometric concepts do not apply to firms"

**Locus.** Not attached to a locus. A frame objection to the paper as a whole.

**What the referee will say.** *"Classical test theory was developed for human respondents answering items. Firms are not respondents and ESG ratings are not test scores. The transfer is unargued."*

**Why it is raised.** It is a category objection with no content. Classical test theory is not a theory about respondents; it is a decomposition of any observed score into a component common to a set of measures and a residual, and it applies wherever several instruments are directed at the same object on a defined scale. The manuscript's own Section 4 shows that the Regulation's definition of an ESG rating, "an established methodology and a defined ranking system of rating categories", is a score in exactly that sense. The credit-rating precedent in the same section does further work.

**Verdict.** Reviewer-noise. Do not concede it. It is worth having a two-sentence answer ready because it is the easiest objection for a referee outside measurement to reach for, and conceding it would cost the paper its whole frame.

**Fix.** Move a short form of the Section 4 warrant forward into Section 2, where the model is first applied, so the reader meets the justification before the arithmetic. Drafted response in Section 5.

**If unfixed.** Nothing, if the answer is ready.

## [13] Reviewer-noise - "The author should have bought current data"

**Locus.** Not attached to a locus.

**What the referee will say.** *"The analysis should have been run on a current common sample."*

**Why it is raised.** The data are proprietary, the cost is substantial, and no single provider's licence would give the common sample the analysis requires. A referee raising this is asking for a different paper. The answer is already in the manuscript's Section 5: the computation is in the gift of a regulator empowered to require every provider's ratings.

**Verdict.** Reviewer-noise. Recorded so it is recognised for what it is if it arrives.

**Fix.** Sharpen the Section 5 ask so the answer is already on the page. Drafted response in Section 5.

**If unfixed.** Nothing.

## 4. Statistics and reporting check

### 4a. Do the numbers support the claims

I recomputed the paper's chain from the published matrix. The arithmetic is sound at every step I checked.

**The reliability estimate.** The mean of the fifteen pairwise correlations is 0.54, and under strict parallelism that is the reliability of a single rating. Correct as classical test theory, correct as arithmetic, and correctly described as "the single-rater form of standardised alpha, rather than a fitted model". The model is rejected by the data, which the manuscript says, and the estimate is retained as a summary, which is legitimate and is flagged. The one-factor alternative is the right second reading and is reported. The interpretive problem is objection 1 and it is not a problem with this quantity; it is a problem with what the complement of this quantity is called downstream.

**The standard error and the intervals.** The square root of one minus 0.541 is 0.678, the minimum significant difference is 1.96 times root two times 0.678, which is 1.88 standard deviations, and on a ten-point scale that is the 13 points and 19 points the manuscript reports. Table 2's bands follow from the standard regression of true score on observed score. Each reproduces. The band for a median firm is 9th to 91st percentile as stated.

**The rank results.** The Pearson correlation between rating and consensus standing is the root of the reliability, 0.736, reported as 0.74. The rank correlation of 0.72 is the correct Spearman transform. The rank standard deviation of 229 positions for the median firm of a thousand and the top-decile hit rate of 50 per cent are simulation outputs and I did not reproduce the simulation; the closed-form quantities they sit beside are right, and the deposited script and seed make them checkable, which is more than most papers offer.

**The Spearman-Brown result.** Six ratings at 0.541 gives 0.876 on the parallel assumption, reported as 0.88. Correct.

**Where the chain does not carry the claim.** Two places, both covered above. The resolution index at objection 2 is the paper's own construction and the established alternative gives roughly half its value. The heterogeneity comparison at objection 4 has no valid null behind it.

**Reliability evidence for an interpreted score.** This is the paper's own subject and it meets its own standard for the object it studies. It does not meet it for its own estimates: see objection 3.

**Measurement comparability.** The paper is about comparability across raters and treats it correctly as the central question. The one gap is the within-industry case at objection 8, where the comparability that matters for the dominant use is unavailable and a pooled figure stands in for it.

### 4b. Reporting completeness and errors

| Item                             | State                                                                                               |
|----------------------------------|-----------------------------------------------------------------------------------------------------|
| Sample size stated and justified | Yes. 924 firms, the common sample of Berg et al.'s Table II, with the reason the year is 2014 given |
| Source of the data stated        | Yes, precisely, including which table and which row                                                 |
| Degrees of freedom reported      | Yes for both chi-squares, 9 and 8                                                                   |

| Item                                   | State                                                                                                                                                                                                |
|----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Exact test statistics                  | Chi-squares given to one decimal. No p-values reported for either                                                                                                                                    |
| Fit indices                            | Missing. No CFI, TLI, RMSEA or SRMR. RMSR of 0.02 given for the second solution only                                                                                                                 |
| Effect sizes                           | Not applicable in the usual sense; the estimands are the effect sizes                                                                                                                                |
| Confidence intervals                   | Missing throughout. No interval on any reliability estimate, communality, omega, hit rate or resolution figure. See objection 3                                                                      |
| Monte Carlo error                      | Missing. 2,000 replications per cell stated; no simulation standard error                                                                                                                            |
| Seed and software                      | Yes. Python 3.13, numpy 2.4, scipy 1.17, seed 20260913. Better than most                                                                                                                             |
| Missing-data handling                  | Not applicable. Complete published matrix                                                                                                                                                            |
| Assumptions stated                     | Yes: normality, common error variance, uncorrelated errors, all named. Only the third is examined, and only for one pair                                                                             |
| Assumptions tested                     | Partially. See objections 5 and 6                                                                                                                                                                    |
| Tables agree with text                 | Yes, with the single exception at objection 9, where the abstract's 0.86 meets the tables' 0.88                                                                                                      |
| Abstract numbers appear in the results | Yes. 0.54, 0.25 to 0.74, 0.86, the 9th-to-91st band and the three-or-four-bands figure all appear in the body with their derivations                                                                 |
| p-value reporting errors               | None. No p-values are reported at all, which is itself objection 3                                                                                                                                   |
| Rounding consistency                   | Consistent to two decimals for correlations, one for chi-squares                                                                                                                                     |
| Table notes                            | Substantive and complete. Each of the three carries its computation, its source and its caveats. Better practice than most submitted manuscripts, and worth keeping intact under any length pressure |

**Four reporting gaps, all in the same family.** No intervals, no p-values, no fit indices, no Monte Carlo error. Individually each is minor. Together they are a pattern, and in this paper the pattern is the thing a referee will name, because the paper's argument is that uncertainty should be reported.

## 5. Response-to-reviewers skeleton

Drafted for the objections judged defensible, plus the two pieces of reviewer-noise likely enough to arrive that an answer should be ready. Complete these in your own words; the register below is deliberately measured, including where it disagrees.

### On [7]: the 2014 data vintage

*The point, as a referee will put it:* "The paper draws conclusions about a 2026 Regulation from a correlation matrix of 2014, in a market that has since consolidated and re-tooled."

**Response:** We agree that the vintage is a real limitation and we have made it more prominent rather than less. The paper's claim is deliberately about what the published evidence implies and not about any provider's current product, and we have moved that statement into the second paragraph of the introduction so that it governs the reader's understanding of everything that follows rather than appearing as a series of caveats. We would resist the implication that the conclusion is therefore unsafe. The 2017 replication reported by the same authors gives a mean of 0.54 against 0.57 for the same five providers in 2014, so the aggregate figure is stable across the one interval for which two

observations exist, while the provider-level pattern moves. More importantly, the paper's argument does not depend on the coefficient being current: it is that a regulation which names reliability as an aim asks for no reliability evidence of any kind, and that argument would be unaffected if the true current coefficient were 0.65.

**Change made:** *[Consolidated the four separate vintage caveats into one statement at the end of the introduction, p. X; strengthened the Section 5 recommendation from a general invitation into a specific proposal naming the data source, the cost and the party best placed to bear it.]*

## On [8]: the within-industry coefficient

*The point, as a referee will put it: "Best-in-class selection operates within industry peer groups, the paper concedes the within-industry coefficient is unpublished, and a random draw of fifty firms from the pooled universe is not a substitute for it."*

**Response:** We accept that the random group of fifty is not a peer group and we have said so more plainly. We would make two points in defence of reporting it. First, the direction of the bias is signable in each of the two cases that arise, and the paper now sets this out in the text rather than in a table note: where both ratings are absolute, industry membership is part of what providers agree about, so the pooled figure is an upper bound on within-industry agreement and the resolution figure we report is optimistic; where one rating is industry-relative and the other absolute, which Annex III point 1(j) anticipates, the pooled correlation can understate it. Second, the unavailability is itself part of the finding. A coefficient that the dominant use of the product depends on is not published by anyone, and that is an instance of the gap the paper documents rather than an evasion of it.

**Change made:** *[Moved the bias argument from the Table 3 note into Section 3 at the point the group-of-fifty figure is first reported, p. X; relabelled the column to make the random-draw basis explicit; added the unavailability to the list of absent quantities in Section 4.]*

## On [12]: whether psychometric concepts apply to firms

*The point, as a referee will put it: "Classical test theory was developed for human respondents answering items. Firms are not respondents and ESG ratings are not test scores."*

**Response:** We respectfully disagree, and we think the Regulation settles the point rather than the measurement literature. Classical test theory is not a theory about respondents; it is a decomposition of any observed score into a component common to a set of measures and a residual, and it applies wherever several instruments are directed at the same object on a defined scale. Article 3(1) of the Regulation defines an ESG rating as an opinion delivered through "an established methodology and a defined ranking system of rating categories", which is a score in precisely the sense the Standards use the term. The paper makes this argument in Section 4 and we have moved a short form of it forward into Section 2, where the model is first applied, so that the reader meets the justification before the arithmetic rather than after it.

**Change made:** *[Added two sentences at the head of Section 2 carrying the Article 3(1) definition forward as the warrant for the measurement reading, p. X.]*

## On [13]: whether current data should have been purchased

*The point, as a referee will put it: "The analysis should have been run on a current common sample."*

**Response:** We would welcome that analysis and cannot perform it. The six providers' ratings are proprietary and separately licensed, and no single licence yields the common sample the computation requires; the 2014 matrix exists in the literature only because Berg, Koelbel and Rigobon assembled it. This is the reason the paper's closing recommendation is addressed to the two parties who can produce it: each provider, which can compute and publish its own reliability and standard error from its own archive at the annual methodology review Article 15(8) already requires, and ESMA, which Article 32 empowers to require the ratings it needs and which must under Article 52(2)(f) assess by 2 January 2029 whether the Regulation has improved the reliability of ESG ratings. We have sharpened that passage so the ask is specific rather than general.

**Change made:** *[Rewrote the final paragraph of Section 5 to name the two parties, the mechanism available to each and the deadline that makes it timely.]*

## On [1]: the error interpretation, if you choose to argue rather than change

*The point, as a referee will put it: "The complement of a cross-provider agreement coefficient is not error variance and cannot be reported as a standard error of measurement."*

**Response:** I do not recommend arguing this one. The change at objection 1 is a renaming that costs the paper nothing and removes its largest exposure. A block is drafted here only because you may meet a referee who has read Section 2 and wants the concession defended even after the revision. The line to take: we agree, and the revision no longer does so. The quantity is now named cross-provider dispersion throughout and defined as the standard deviation of the placements a firm of given consensus standing receives across providers of equal typicality. We retain the term standard error of measurement only in Section 4, where it denotes the quantity the Standards require a provider to report about its own instrument, which is a different quantity and is bounded above by ours. Distinguishing the two is now one of the paper's contributions rather than a source of ambiguity in it.

**Change made:** *[Renamed throughout Sections 3 and 5 and in the abstract; "true standing" replaced by "consensus standing" in the Table 2 stub and note, the Table 3 hit-rate definition, and three occurrences in Section 3.]*

## 6. Readiness statement

This manuscript is closer to ready than its current destination suggests, and the gap between those two facts is the most useful thing in this report.

The analysis is correct and unusually well documented. Its table notes carry their own derivations, the script and seed are deposited, and the argument in Section 4 is the part of the paper I would not change: setting the Regulation's seventeen disclosure items beside what a validity argument requires is a genuine contribution and it is made carefully. Four of the twelve objections above are real defects in reporting rather than in substance, and three of those four are a week's work between them.

The largest remaining risk is not any single objection. It is that the paper arrives at a journal that will not read it. At Journal of Business Ethics the anonymisation failures alone are a probable return before review, the length sits 2,500 words below the stated range for the article type, no business-ethics literature is engaged, and the SSRN posting the author intends is in direct conflict with the journal's pre-publication rule. Fixing the mechanical rows in Section 2 would not fix the scope row, and the scope row is the one that decides this.

What I would do first, in order. Decide the venue before touching the manuscript, because the answer changes the length target, the literature to engage and whether the SSRN posting happens at all. On the evidence of the paper itself, its natural homes are a measurement or methodology venue, or a sustainable-finance venue that publishes quantitative methods work, and at either of those the scope objection disappears and the contribution reads as the contribution rather than as a response to Berg et al. Then rename the quantity at objection 1, which is a day and removes the largest substantive exposure. Then bootstrap the intervals at objection 3 and add the strata comparison at objection 2, which together are the difference between a paper that anticipates its reviewers and one that is caught by them.

**This is not a prediction of acceptance and no part of this report should be read as one. What a journal decides depends on the editor, on which reviewers are drawn, and on the fit of the paper to that issue, none of which is within my knowledge. A manuscript reviewed here can be rejected, including for reasons I have identified and you choose not to address, and including for reasons I have not anticipated.**

One thing this report does not do, and should not be read as doing: it does not establish that the underlying analysis is sound beyond the checks named in Section 4a. I recomputed the closed-form quantities from the published matrix. I did not re-run the simulation, and nothing here is a research-integrity assessment.

## 7. Signature and standing statements

**Dr. Milos Kankaras**

Psychometrician. ORCID 0000-0002-3190-7751

24 September 2026

### The commitments this report is written under

**This is advice to the author. It is never co-authorship and never ghost-writing.** The report is written about the manuscript and not into it. No sentence in it is intended to be pasted into the paper. Suggested fixes describe the change that is needed; they do not supply the text that makes it. The reviewer acquires no authorship claim and no interest of any kind in the work.

**No outcome is promised and no acceptance is implied.** This report says what a referee will probably object to. It does not say what an editor will decide. No report, page or message from this service carries an acceptance probability, a readiness score or a grade.

**Recusal.** Having been paid by an author, the reviewer will never afterwards review that manuscript for anybody else. Every engagement is entered in the AdriaMont conflict register at intake, before the report is written, with the manuscript title, the author, the target venue and the engagement

date. Recusal from the reviewed manuscript is permanent and covers any reviewing capacity: journal referee, editor, funder or evaluation panel, prize jury, thesis examiner, promotion or tenure assessor, including later versions under a changed title. For twenty-four months the paid relationship is disclosed before any review of other work by the same author.

## How your manuscript is handled

Your manuscript is confidential. It is stored in the folder for your engagement and nowhere else. It is never used as training data, never added to any corpus or example library, and no part of it appears in any sample, page or talk. It is not forwarded to anyone; there is no second reader. It is deleted on your request, at any time, and the deletion is confirmed to you in writing. Otherwise it is deleted six months after the report is delivered.

## This service

|                                |                                                                                                                                                                                      |
|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Tier 1, report                 | 550 EUR ex VAT (385 EUR for doctoral candidates)                                                                                                                                     |
| Tier 2, report and one re-read | 850 EUR ex VAT (595 EUR for doctoral candidates)                                                                                                                                     |
| Turnaround                     | Five working days for the report, three for the re-read                                                                                                                              |
| Scope                          | Quantitative social science, measurement and psychometrics, psychology, education and educational assessment. Manuscripts outside those fields are declined and the fee is not taken |

AdriaMont Consulting, Vijenci Danila Kisa 5, 81000 Podgorica, Montenegro. PIB 03563227. CRPS 5-1146711/005. VAT-registered.

**Dr Milos Kankaras, Co-founder, AdriaMont Consulting** | [milos@adriamont.me](mailto:milos@adriamont.me) | [miloskankaras.com](https://miloskankaras.com) | [adriamont.me](https://adriamont.me) | [ORCID](#)