

Two Numbers, One Record: A Study Protocol for Testing Whether Separate Reporting of Planned and Extra Output Makes IMF Records Comparable as Creative-Campus Data

Untung Rahardja*, Qurotul Aini, Nuke Puji Lestari Santoso,
Marviola Hardini, Fitra Putri Oganda

Universitas Raharja, Indonesia

*Corresponding author: urahardja@esquine.com

Q. Aini: aini@raharja.info; N. P. L. Santoso: nuke@raharja.info

Abstract

Indicator Measurement Factor Analysis (IMF) records creative-campus projects as missions with planned outputs, extra outputs and an achievement percentage. A concept analysis earlier in this series proposed (P4) that reporting planned achievement and extra output as two separate numbers makes IMF records comparable across missions and cohorts and therefore usable as campus-level creative-economy data. This paper is a conceptual and study-protocol paper that prepares the test of that proposition; it reports no empirical results. We show that a single combined percentage admits at least three defensible readings once a mission produces extra output, and model a reported figure as a record component plus a reader component with a judgement part and a convention part, the latter removed by two-number reporting. We state four hypotheses with pre-specified decision rules, and give a fully crossed rater protocol in which independent, trained raters recompute figures from the same de-identified records under a single-percentage condition, a two-number condition and a sensitivity condition with a fully specified single formula. The analysis plan uses the absolute-agreement intraclass correlation ICC(A,1) and Krippendorff's alpha with paired bootstrap comparisons, limits of agreement, a mixed model for moderation by the share of extra output, robustness checks and a planning simulation for sample size (80 records, four raters). A worked example applying the protocol to six missions shows the computations end to end and how the convention part alone can lower agreement on a single figure without any rater error. The protocol is designed so that support for the two-number convention, support for a precisely specified single rule, or a negative result would each be informative for campuses that report creative-economy outputs.

Keywords: Indicator Measurement Factor Analysis; IMF; inter-rater agreement; intraclass correlation; Krippendorff's alpha; study protocol; output reporting; creative campus; creative economy; project records

1. Introduction

1.1 A Percentage That Means Different Things

Creative-campus projects produce outputs that are easy to admire and hard to count. A student team finishes a prototype, registers a copyright, submits a paper and, along the way, also runs a workshop that nobody planned. A lecturer's incubation cohort delivers most of what it promised and several things it did not promise. When a faculty, a university or a ministry programme later asks how much was achieved, the answer is usually a single percentage. Yet the same project record can yield different percentages depending on who reads it. One reader adds the unplanned outputs to the numerator and reports more than 100 percent. Another caps the figure at 100 percent. A third adds the unplanned outputs to both numerator and denominator and reports

less than 100 percent. All three readers are honest; they simply apply different conventions to a number that hides two quantities.

This matters because the creative economy increasingly asks universities for data. Monitoring platforms for grants and research projects promise real-time tracking, structured reporting and transparency [1, 2], dashboards summarise the performance of staff and units [3], and project-control methods such as earned value management rest on the assumption that progress figures are computed in the same way everywhere [4]. Studies of incubation and creative entrepreneurship rely on comparable output measures to identify what makes programmes succeed [5, 6], and work on monitoring for organisational goals stresses that measurement only helps management when its meaning is shared [7]. A percentage whose meaning depends on the reader is not data in this sense, however carefully it was entered.

1.2 IMF and Its Fourth Proposition

Indicator Measurement Factor Analysis (IMF) is a spreadsheet-based method for running projects as bounded missions. It was introduced to a scientific audience in a 2020 conference paper that describes a Viewboard header and a Content grid in which each contributor completes coded tasks with a percentage achievement [8]. In practice, IMF sheets record planned outputs (Luaran Target), extra outputs (Luaran Extra) and their sum (Luaran Total), together with an achievement figure (Capaian) [9, 10]. A concept analysis earlier in this working-paper series redefined IMF, treated achievement as a bounded ratio with extra output reported separately, and derived five testable propositions [9]. Its fourth proposition, P4, reads as follows.

P4. Reporting planned achievement and extra output separately makes IMF records comparable across missions and cohorts and therefore usable as campus-level creative-economy data.

The concept analysis named the indicator, the agreement between achievement figures recomputed by independent raters from the same records, and the test, a comparison of inter-rater agreement under the two-number convention and under a single combined percentage. A companion paper turned IMF into a process framework for creative-campus projects and adopted the two-number rule as one of its records rules [10], and a third paper proposed an indicator set and maturity model in which output measures play a central part [11]. None of these papers tested P4. This paper prepares that test.

1.3 Aim and Contribution

The aim of this paper is to specify, before any data are collected, how P4 should be tested. It is a conceptual and study-protocol paper. It reports no empirical results, and no IMF records or rater data were analysed for it. Its contributions are four. First, it develops the measurement argument behind P4: why a single combined percentage invites disagreement, and why two numbers should not. Second, it states P4 as a set of directional hypotheses with explicit decision rules. Third, it gives a complete rater protocol, from the sampling of records to the blinding of raters and the handling of missing evidence. Fourth, it specifies the agreement statistics, the intraclass correlation coefficient (ICC) and Krippendorff's alpha, together with the planned comparisons, sample-size reasoning and robustness checks, and demonstrates them with a small worked example that applies the protocol end to end.

1.4 Structure of the Paper

Section 2 reviews the IMF record and the measurement problem. Section 3 develops the framework and states the hypotheses. Section 4 sets out the study design and the rater protocol. Section 5 specifies the agreement statistics and the analysis plan. Section 6 presents a worked example of the protocol. Section 7 discusses validity threats, Section 8 implications and limitations, and Section 9 concludes.

2. The IMF Record and the Measurement Problem

2.1 What an IMF Record Contains

The 2020 conference paper describes an IMF as a Viewboard, a header that displays the project title, objectives, outcomes, contributors, version and due date, and a Content grid that holds the coded tasks [8]. The Viewboard idea itself grew out of earlier work on compact information displays in the same community [12]. Practice sheets examined in the concept analysis and in the operating-system paper add a countdown, a closing flag, owners for each task (PIC) and for each group of tasks (LPIC), evidence links, and three output counts: Luaran Target, the number of coded tasks in the planned section; Luaran Extra, the number of tasks in a separate extra section; and Luaran Total, their sum [9, 10]. The achievement cell, Capaian, is computed from task achievement, but the sheets differ in what they divide by: some use the target, others the total.

For the purposes of this paper an IMF record is the set of fields from which an achievement figure can be computed by someone who did not keep the sheet. Table 1 lists those fields. The record is the unit of analysis in the protocol of Section 4.

Table 1. Fields of an IMF record used to compute achievement

Field	Content in the sheet	Use in this protocol
Purpose and dates	Purpose statement, start date, due date, countdown, closing flag	Eligibility (closed missions only); stratification by mission type
Planned section	Coded tasks counted as Luaran Target (T)	Denominator of planned achievement
Extra section	Coded tasks added during the mission, counted as Luaran Extra	Source of extra output (E)
Task state	Checkbox or achievement value per row	Completion judgement (P and E)
Evidence link	Link to the artefact for each task	Evidence judgement; missing evidence counts as not completed
Owners	PIC and LPIC per task	Removed or replaced by role labels before rating
Achievement cell	Capaian, computed by the sheet	Hidden from raters; kept for comparison only

2.2 Three Readings of One Percentage

When planned and extra outputs are merged into one figure, at least three conventions are defensible, and each appears in ordinary reporting. Let T be the number of planned outputs, P the number of planned outputs completed with evidence, and E the number of extra outputs completed with evidence. Reading A, the uncapped ratio, reports $(P + E) / T$ and can exceed 100 percent. Reading B, the capped ratio, reports $\min(P + E, T) / T$ and never exceeds 100 percent. Reading C, the total-based ratio, reports $(P + E) / (T + E)$, which treats extra outputs as if they had been planned. A fourth variant, which counts partial task achievement rather than completed tasks, multiplies the possibilities.

The readings agree only when $E = 0$. As soon as a mission produces extra output, which is common and, in creative work, desirable [13, 14], the same record yields different numbers. Reading A rewards extra output but makes figures above 100 percent hard to compare with figures below it. Reading B hides extra output entirely, so a team that did more than planned looks identical to one that did exactly what was planned. Reading C can make a team that did more look as if it did less: a mission with eight planned outputs, eight completed and two extra reports 100 percent under reading C, while the same mission without the extra outputs also reports 100 percent, and a mission with six of eight planned outputs plus two extra reports 80 percent, although it fell short of its plan by a quarter.

2.3 The Two-Number Convention

The two-number convention reports planned achievement, P / T , which is bounded between 0 and 100 percent, and extra output, E , as a count, separately. It was proposed in the concept analysis as part of the redefinition of IMF [9] and adopted in the operating-system paper as a records rule [10]. It has two intended properties. It removes the convention choice from the reader, because neither number requires a decision about how to combine the two quantities. And it keeps extra output visible without letting it inflate or deflate planned achievement, which matters for teams whose motivation depends on recognition of extra effort [15, 16].

The convention does not remove every source of disagreement. Raters can still disagree about whether a task was completed, whether its evidence link is valid, and whether a task belongs to the planned or the extra section. P4 therefore does not claim perfect agreement under the two-number convention. It claims that agreement will be higher than under a single figure, because one source of disagreement, the combination rule, has been removed while the others remain common to both conditions.

2.4 Why Agreement Is the Right Test

Comparability across missions and cohorts has several components: the same fields must exist, they must be filled in the same way, and the figures computed from them must mean the same thing. The first two are matters of template design and training, addressed by the operating-system and indicator papers [10, 11]. The third is a matter of measurement reliability. If independent raters, given the same record and the same written rule, compute the same figure, then the figure is a property of the record and not of the reader, and figures from different records can be compared. If they do not, any campus-level aggregate inherits the disagreement. Inter-rater agreement is therefore a necessary condition for P4, and a direct test of the mechanism the proposition names. It is not a sufficient condition for usefulness as creative-economy data, a point we return to in Section 7.

3. Framework and Hypotheses

3.1 A Measurement Model of Reported Achievement

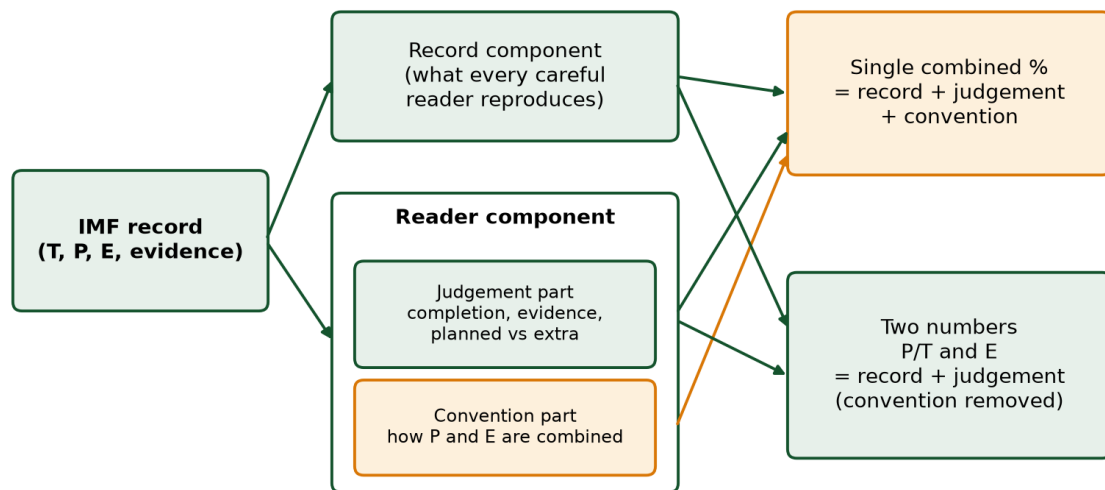
We model a reported achievement figure as the sum of a record component, which every careful reader would reproduce, and a reader component, which depends on the person computing it. The reader component has two parts. The judgement part arises from decisions about individual tasks: completion, evidence and classification as planned or extra. The convention part arises from the rule used to combine the counts into one figure. Under a single combined percentage both parts are present. Under the two-number convention the convention part is, by construction, absent, provided the rule is written down. Figure 1 summarises the model.

The model has an implication that guides the design. The convention part only exists when extra output exists. Missions with $E = 0$ should show similar agreement under both conditions, and the advantage of the two-number convention should grow with the share of extra output. This gives the test a dose-response structure that a simple comparison of means would miss.

3.2 Hypotheses

We state P4 as four hypotheses. H1 and H2 are confirmatory; H3 and H4 are secondary.

H1. Absolute agreement among independent raters on planned achievement (P / T) under the two-number convention is higher than absolute agreement on the single combined percentage computed from the same records. Measure: ICC for absolute agreement, single rater, two-way random effects, ICC(A,1), with Krippendorff's interval alpha as a co-primary statistic.



Model: the convention part exists only when $E > 0$, so the advantage of two numbers should grow with the share of extra output (H3).

Figure 1. Measurement model of a reported achievement figure: the convention part of the reader component is present in a single combined percentage and removed by two-number reporting

- H2.** Agreement on extra output (E) under the two-number convention reaches at least the level conventionally described as good, ICC(A,1) of 0.75 or more [17], with the lower bound of its 95 percent confidence interval above 0.60.
- H3.** The difference in agreement between the two conditions increases with the share of extra output in the record, $E / (T + E)$.
- H4.** Under the two-number convention, the remaining disagreement is concentrated in task classification (planned versus extra) and evidence judgement, not in arithmetic, as shown by the coded reasons raters give for their figures.

3.3 Decision Rules

H1 is supported if the difference in ICC(A,1) between the two-number and single-figure conditions is positive and its 95 percent bootstrap confidence interval excludes zero, and if Krippendorff's alpha points in the same direction. It is not supported if the interval includes zero, and it is contradicted if the interval lies wholly below zero. H2 is supported if both criteria in its statement are met. H3 is supported if the interaction term described in Section 5.5 is positive with a 95 percent interval excluding zero. H4 is descriptive and is reported as proportions with confidence intervals. A practical threshold accompanies H1: we regard a difference in ICC of 0.10 or more as practically meaningful, because it moves a measure across one of the conventional bands of reliability [17]. All decision rules are fixed before data collection and should be registered with the protocol.

Table 2. Hypotheses, measures and decision rules

Hypothesis	Measure	Supported if	Status
H1: higher agreement on P/T under two numbers than on a single figure	ICC(A,1) and Krippendorff's interval alpha, paired bootstrap difference	Difference > 0 with 95% CI excluding 0 for ICC; alpha in same direction	Confirmatory
H2: good agreement on extra output E	ICC(A,1) for E	Point estimate ≥ 0.75 and lower 95% bound > 0.60	Confirmatory

Hypothesis	Measure	Supported if	Status
H3: the advantage grows with the share of extra output	Condition $\times$ share interaction in a mixed model of absolute deviations	Interaction > 0 with 95% CI excluding 0	Secondary
H4: remaining disagreement lies in classification and evidence	Proportions of reason codes on discrepant tasks	Described with 95% CIs; no test	Secondary

4. Study Design and Rater Protocol

4.1 Design Overview

The study is a fully crossed reliability design with a within-record manipulation. Every rater computes figures from every record under both reporting conditions, so that the comparison is not confounded by differences between records or between raters. Records are drawn from IMF sheets of completed missions; raters are trained readers who did not keep the sheets. The two conditions differ only in the reporting instruction: in the single-figure condition, raters are asked for the mission's achievement as one percentage; in the two-number condition, they are asked for planned achievement and extra output separately. Figure 2 shows the flow.

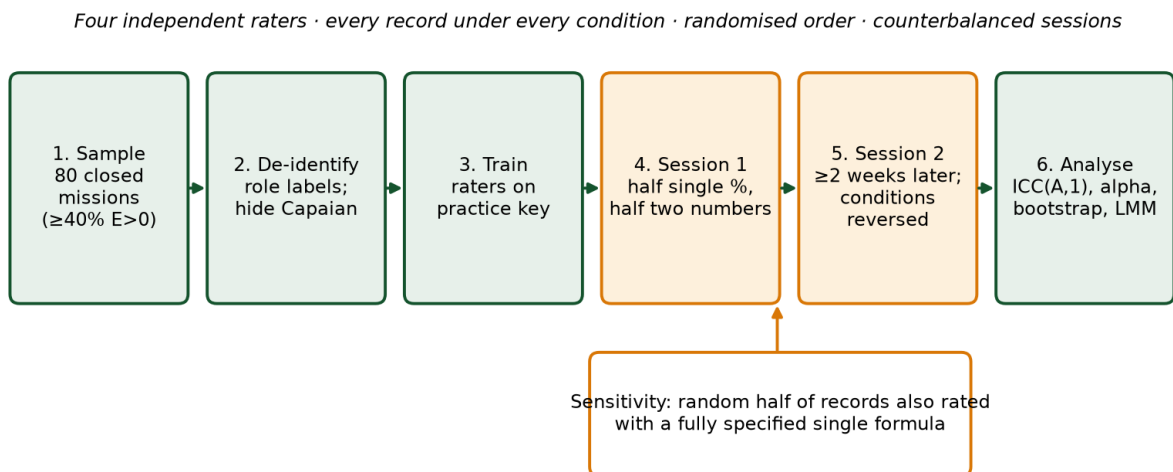


Figure 2. Flow of the proposed rater study

4.2 The Record Corpus

The sampling frame is the set of closed IMF missions in a defined period at one or more universities, for example all missions closed in one academic year in programmes that use IMF for creative-campus work, such as incubation, competitions, publication projects and community events [6, 18, 19]. A mission is eligible if its sheet has a planned section, a due date and at least three coded tasks. Missions are stratified by type of output and by whether the extra section contains any completed task, and drawn so that at least 40 percent of the sample has E greater than zero; without such records the comparison has nothing to detect. Records are exported as read-only copies. Names of contributors are replaced with role labels before raters see them, and recognition cells that name individuals are removed, because they are not needed to compute achievement and could bias judgement.

4.3 Raters and Training

Raters are recruited from staff and advanced students who have used spreadsheets for project work but have not kept any of the sampled sheets. We plan four raters, a number that allows a fully crossed design within a feasible workload while giving the variance components a reasonable basis [20]. Training uses a separate set of practice records, not drawn from the sample, and covers the record fields, the definition of completion and evidence, and the written rule for each condition. Training ends when each rater reproduces the figures of a key on five practice records. The written rule for the single-figure condition deliberately reproduces the instruction found in practice, “report the mission’s achievement as a percentage”, without specifying how to treat extra output; this is the condition whose reliability P4 questions. A sensitivity condition with a fully specified single-figure rule is described in Section 4.6.

4.4 Procedure and Blinding

Each rater receives the records in an individually randomised order, in two sessions separated by at least two weeks. In one session the rater applies the single-figure instruction to half of the records and the two-number instruction to the other half; in the second session the assignment is reversed. The order of conditions is counterbalanced across raters. The separation and counterbalancing limit the chance that a rater remembers the earlier figure for the same record. Raters work independently, do not see each other’s figures and do not see the figure in the sheet’s own Capaian cell, which is hidden in the exported copy. For each record a rater enters the figures requested, the counts T, P and E as they read them, and, for every task whose status they found unclear, a reason code from a short list: completion unclear, evidence missing or broken, classification unclear, other.

4.5 Handling Missing and Ambiguous Evidence

A task counts as completed only if its row shows completion and its evidence link opens to an artefact that matches the task. Broken links, links to empty folders and links that require permissions the rater does not have are coded as missing evidence and the task is counted as not completed. This rule is applied identically in both conditions, so it affects the judgement part of the reader component equally and does not favour either condition. Records in which more than half of the tasks lack working evidence are retained, because excluding them would make the corpus look more reliable than practice is, but they are flagged for the robustness check in Section 5.6.

4.6 Sensitivity Condition

A critic could argue that the single-figure condition is unfairly underspecified, and that a precisely written single rule, for example “report $(P + E) / T$, uncapped”, would perform as well as two numbers. We therefore add a third condition for a random half of the records: a single percentage with a fully specified formula. If agreement in this condition matches the two-number condition, the benefit of P4 lies in specifying the rule rather than in separating the numbers, and the conclusion should say so. If agreement remains lower, or if the specified single figure is reliable but loses the information about extra output, the two-number convention is the better design. This condition protects the study from confirming its own framing.

Table 3. The rater protocol in brief

Step	What happens	Safeguard
1. Sampling	80 closed missions, stratified by type; at least 40% with $E > 0$	Pre-specified eligibility; read-only export

Step	What happens	Safeguard
2. De-identification	Names replaced with role labels; recognition and Capaian cells hidden	Key stored separately
3. Training	Practice records outside the sample; key to be matched on five records	Same training for all raters
4. Session 1	Half the records single-figure, half two-number; random order per rater	Independent work; counterbalanced order
5. Session 2 (≥ 2 weeks later)	Conditions reversed for each record	Limits memory effects
6. Sensitivity condition	Random half of records with a fully specified single formula	Tests whether rule specification alone suffices
7. Reason codes	Completion, evidence, classification or other for every unclear task	Supports H4

4.7 Ethics and Data Protection

The study analyses records of project work, not people. Even so, records contain names and links to student work. The protocol therefore requires approval from the institution’s research ethics committee or equivalent body, permission from the owners of the sheets, de-identification before rating, read-only copies, and storage of the rating data separately from any key that links records to teams. Raters are told that the study evaluates reporting conventions, not their skill, and their individual figures are reported only in aggregate. Because the first author is the originator of the IMF method, data collection and analysis should be carried out by team members without a stake in the outcome, and the analysis script should be fixed before the data are unblinded.

5. Agreement Statistics and Analysis Plan

5.1 Intraclass Correlation

The intraclass correlation coefficient expresses the share of variance in ratings that is due to differences between records rather than to raters and error [20,21]. Because the question is whether raters produce the same number, not merely numbers in the same order, the primary form is the absolute-agreement ICC for a two-way random-effects model with a single rater as the unit, ICC(A,1) in the notation of McGraw and Wong [21], corresponding to ICC(2,1) in that of Shrout and Fleiss [20]. With n records rated by k raters, and mean squares for records (MSR), raters (MSC) and residual error (MSE) from a two-way analysis of variance, it is

$$ICC(A,1) = \frac{MSR - MSE}{MSR + (k-1)MSE + \frac{k}{n}(MSC - MSE)}$$

The random-effects choice treats the raters as a sample from the population of people who might read IMF records, which is the population to which campus-level data would be exposed. We report ICC(A,1) with its 95 percent confidence interval and interpret it with the bands of Koo and Li: below 0.50 poor, 0.50 to 0.75 moderate, 0.75 to 0.90 good and above 0.90 excellent, applied to the confidence interval rather than to the point estimate [17]. The consistency form ICC(C,1) is reported alongside it, because a large gap between the two indicates a systematic difference between raters, which is exactly what a convention difference produces.

5.2 Krippendorff’s Alpha

Krippendorff’s alpha compares observed disagreement with the disagreement expected by chance, given the distribution of all values, and is defined as one minus their ratio [22]. It accommodates any number of raters, missing ratings and different levels of measurement. We use the interval metric for planned achievement, the single figure and extra output. We also compute the ratio

metric for extra output as a sensitivity check, because a count has a true zero. Alpha is reported with a bootstrap confidence interval from 10,000 resamples of records, following Hayes and Krippendorff [22]. Values of 0.80 or above are conventionally taken as reliable and values from 0.667 to 0.80 as adequate for tentative conclusions [22]. We report both ICC and alpha because they answer slightly different questions, and agreement between them strengthens any conclusion.

5.3 Agreement in Units of the Figure

Coefficients are relative to the spread of records: a sample of very similar missions can show low reliability even when raters differ by only a few points. We therefore also report agreement in the units of the figures themselves, using the limits of agreement of Bland and Altman for each pair of raters [23], the mean absolute difference between raters, and the proportion of records on which all raters agree within five percentage points for achievement and exactly for extra output. These figures speak directly to a programme manager who wants to know how far apart two readers of the same sheet could be.

5.4 Comparing the Conditions

Because the conditions share records and raters, the ICC values are dependent. The comparison of H1 uses a paired bootstrap: records are resampled with replacement, both ICCs and both alphas are recomputed in each resample, and the 95 percent percentile interval of the difference is taken. The same procedure gives the difference in alpha. Rater-level resampling is added as a check, although with four raters it is coarse.

5.5 Moderation by Extra Output

H3 is tested at the level of rater-record pairs. For each record and condition we compute the absolute deviation of each rater's figure from the record median, and fit a linear mixed model with deviation as outcome, condition, the share of extra output and their interaction as fixed effects, and random intercepts for records and raters. H3 predicts a positive interaction: deviations in the single-figure condition rise with the share of extra output, deviations in the two-number condition do not. For comparability, single-figure and planned-achievement deviations are both expressed in percentage points.

5.6 Robustness and Sample Size

Four robustness checks are planned: excluding records flagged for missing evidence; excluding the rater with the lowest agreement with the training key; computing the single figure only from records with E greater than zero; and replacing the interval metric with the ratio metric for extra output. Conclusions are stated as robust only if they hold in all four.

For sample size, we follow the confidence-interval approach recommended for reliability studies: the number of records should be large enough that the 95 percent interval of ICC(A,1) for the two-number condition is narrow enough to lie within one of the Koo and Li bands [17]. A small planning simulation, which assumes a true ICC(A,1) of 0.85 and four raters and is not a result, gives an expected 95 percent interval half-width of about 0.07 with 40 records, 0.06 with 60 and 0.05 with 80. We therefore plan 80 records, of which at least 32 have extra output, and recommend that the final number be confirmed by repeating the simulation with the variance estimates of the study's own pilot. Table 4 summarises the plan.

Table 4. Analysis plan

Question	Statistic	Reporting
Do raters reproduce the same number?	ICC(A,1), two-way random effects; ICC(C,1) alongside	Estimate and 95% CI; bands of Koo and Li applied to the CI
Is agreement above chance given the value distribution?	Krippendorff's alpha, interval metric (ratio metric for E as check)	Estimate and bootstrap 95% CI (10,000 resamples)
How far apart are raters in units of the figure?	Bland-Altman limits of agreement; mean absolute difference; share within 5 points	Per rater pair and pooled
Does the two-number condition agree more?	Paired bootstrap of ICC and alpha differences	Difference and 95% CI
Does extra output drive disagreement?	Linear mixed model with random intercepts for records and raters	Interaction estimate and 95% CI
Are the conclusions robust?	Four pre-specified exclusions and metric changes	Conclusion stated as robust only if all four agree

6. Worked Example: Applying the Protocol

6.1 Purpose of the Worked Example

This section applies the IMF two-number protocol step by step to a small set of six missions and three raters. The six missions are a worked demonstration of the procedure, not field data collected for this study. The demonstration has two purposes: to show the computations end to end, and to show how the convention part of the reader component behaves when extra output is present.

6.2 The Demonstration Records

Table 5 lists six missions with their planned outputs (T), completed planned outputs (P) and completed extra outputs (E). Three raters compute a single percentage. Each applies a different defensible reading: rater 1 the uncapped ratio (reading A), rater 2 the capped ratio (reading B) and rater 3 the total-based ratio (reading C). Under the two-number convention all three raters compute P / T and E , but rater 3 is assumed to classify one extra task of mission M4 as planned, a judgement slip of the kind the protocol expects to remain.

Table 5. Worked example: six missions and the figures of three raters under the two reporting conditions

Mission	T	P	E	Single % (R1 / R2 / R3)	Planned % (R1 / R2 / R3)	Extra (R1 / R2 / R3)
M1	8	8	2	125.0 / 100.0 / 100.0	100 / 100 / 100	2 / 2 / 2
M2	10	7	0	70.0 / 70.0 / 70.0	70 / 70 / 70	0 / 0 / 0
M3	6	6	3	150.0 / 100.0 / 100.0	100 / 100 / 100	3 / 3 / 3
M4	12	9	4	108.3 / 100.0 / 81.2	75 / 75 / 83.3	4 / 4 / 3
M5	5	4	1	100.0 / 100.0 / 83.3	80 / 80 / 80	1 / 1 / 1
M6	9	9	0	100.0 / 100.0 / 100.0	100 / 100 / 100	0 / 0 / 0

6.3 The Computations

The single-figure ratings disagree whenever E is greater than zero. For mission M3, with six planned outputs, all six completed and three extra, the three raters report 150, 100 and 100 percent; for M4 they report 108.3, 100 and 81.2 percent. Mission M2, which has no extra output, receives 70 percent from every rater, as the model in Section 3.1 predicts. Under the two-number

convention the only disagreement is the slip on M4: planned achievement of 75, 75 and 83.3 percent, and extra output of 4, 4 and 3.

Applying the formulas of Section 5 to these records gives, for the single figure, $ICC(A,1) = 0.48$ and Krippendorff's interval $\alpha = 0.41$, with a largest disagreement of 50 percentage points on one record. For planned achievement under the two-number convention the values are $ICC(A,1) = 0.98$ and $\alpha = 0.98$, with a largest disagreement of 8.3 points; for extra output they are $ICC(A,1) = 0.98$ and $\alpha = 0.97$, with a largest disagreement of one output. The consistency form for the single figure, $ICC(C,1) = 0.59$, is noticeably higher than the absolute form, which is the signature of a systematic rater difference: rater 1's figures are consistently higher than those of rater 3. We checked the values by computing them independently with two open-source implementations. Table 6 sets them out.

Table 6. Worked example: agreement statistics by reporting condition

Figure	ICC(A,1)	ICC(C,1)	Krippendorff's alpha (interval)	Largest disagreement
Single combined %	0.48	0.59	0.41	50 points (M3)
Planned achievement, P/T	0.98	0.98	0.98	8.3 points (M4)
Extra output, E	0.98	0.98	0.97	1 output (M4)

6.4 What the Worked Example Demonstrates

The worked example demonstrates that the convention part of the reader component can by itself lower agreement on the single figure from near-perfect to poor, without any rater making an error, and that it disappears when the numbers are reported separately. It also shows why the moderation test of H3 matters: in the example all of the single-figure disagreement comes from the four missions with extra output. Whether raters in the field behave in this way is the question the full study answers. Real raters may share a convention, in which case the single figure would be more reliable than in the example; or they may disagree about completion and evidence far more often, in which case both conditions would be less reliable. With six records, moreover, any confidence interval would be too wide to support a conclusion. Only the study specified in Sections 4 and 5 can answer the question.

7. Threats to Validity

Construct validity is the first concern. Agreement shows that a figure belongs to the record; it does not show that the record reflects what the team actually produced. Two raters can agree perfectly on an inflated sheet. The protocol addresses this in part through the evidence rule, and the indicator set proposed earlier in this series provides external criteria, such as verified outputs per contributor-month, against which agreed figures could later be validated [11].

Internal validity is threatened by memory and learning. A rater who has computed a single figure for a record may recall it when computing two numbers for the same record. Session separation, randomised order and counterbalancing reduce but do not remove this risk; the comparison by session order, which the counterbalanced design permits, will show whether it mattered.

External validity is limited by the corpus. Records from one university network, where IMF originated, may be more uniform than records from campuses that adopt it later, and raters trained for the study may be more careful than the staff who would read records in routine reporting. The protocol should therefore be replicated at a second institution, with raters drawn from the administrative staff who prepare reports, before the convention is recommended for national creative-economy data [24, 25].

Statistical conclusion validity depends on the spread of records. If the sampled missions are very similar in achievement, coefficients will be low in both conditions even when raters agree closely in absolute terms. Reporting agreement in units of the figure (Section 5.3) and stratified sampling (Section 4.2) guard against misreading such a result.

Finally, there is a risk of investigator allegiance. The first author originated IMF, and the authors have proposed the two-number convention in earlier papers of this series [9, 10]. The sensitivity condition of Section 4.6, the pre-registered decision rules and the separation of data collection from the authors with a stake in the outcome are intended to make a negative result as publishable as a positive one.

8. Discussion

8.1 Contributions

The paper makes three contributions. It turns a proposition stated in one sentence into a test that others can run, with hypotheses, decision rules and an analysis plan fixed in advance. It gives the measurement argument behind the two-number convention a simple model, separating the judgement and convention parts of the reader component, which explains when the convention should help and when it should not. And it connects a local project method to the wider question of how universities produce creative-economy data that can be trusted and compared, a question that monitoring platforms, dashboards and incubation studies have so far approached through systems rather than through the reliability of the figures they display [1, 3, 5].

8.2 Implications for Practice

Three implications follow even before the study is run, because they rest on the logic of Section 2 rather than on any result. First, any campus that reports IMF achievement as a single figure should write down the combination rule next to the figure; without it, the figure cannot be read. Second, extra output deserves its own column in programme reports, because the capped convention silently erases it and the total-based convention penalises it. Third, the evidence rule matters in either convention: a figure that rests on broken links is not more reliable for being split in two. Publication-management and journal-management work in the same community has made similar points about the visibility and verification of outputs [19, 26].

8.3 Limitations

The paper has four limitations. It reports no data, and every statement about agreement under real conditions remains a hypothesis. Its sample-size reasoning rests on approximations and an assumed ICC that a pilot may revise. Its model of the reader component is deliberately simple and ignores, for example, partial task achievement and the weighting of outputs of different size, which some sheets use. And the study it proposes tests reliability, which is only one condition for records to serve as creative-economy data; validity, coverage and the willingness of teams to keep records honestly are others.

8.4 Future Research

The first step is a pilot with ten to fifteen records and two raters to test the procedure, the reason codes and the time required, followed by the main study. Further work could extend the design to partial achievement and weighted outputs, compare IMF records with other project-reporting formats used in campus incubators and learning factories [27, 28], and examine whether separate reporting of extra output changes what teams choose to do, a behavioural question related to the third proposition of the concept analysis [9].

9. Conclusion

A single achievement percentage looks like data, but when a mission produces more than it planned, the percentage depends on who computes it. The fourth proposition of the IMF concept analysis holds that reporting planned achievement and extra output as two separate numbers removes this dependence and makes IMF records comparable as campus creative-economy data. This paper has given that proposition a measurement model, four hypotheses with decision rules, a crossed rater protocol with a sensitivity condition, and an analysis plan built on the intraclass correlation and Krippendorff's alpha, demonstrated with a worked example that applies the protocol end to end. The study it describes can support the two-number convention, show that a precisely written single rule is enough, or show that neither is reliable without better evidence practices. Each outcome would tell campuses something useful about the figures they report.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest. Untung Rahardja is the originator of the IMF method and a co-author of the 2020 conference paper that introduced it; the other authors are colleagues at the same university. The study protocol includes safeguards against investigator allegiance (Sections 4.6, 4.7 and 7). The authors declare no other conflict of interest.

Data availability. No data were collected or analysed for this paper. The worked example in Section 6 applies the protocol to demonstration records; the short script that computes its coefficients is available from the corresponding author on request.

Use of AI tools. Generative AI tools were used to assist with language editing, drafting support and formatting of this manuscript. The authors reviewed and verified all content, computations and references and take full responsibility for the work.

References

- [1] U. Rahardja, L. Sulisty, D. Safarina, M. Rapidan Kusuma, N. Silawati, and Z. Nanle, "HibahQu Education Monitoring Platform Based on Human-Centric Orange Technology Laravel 12 Vue.js," *ADI Bisnis Digit. Interdisiplin J.*, vol. 6, no. 2, pp. 203–218, Dec. 2025, doi: [10.34306/abdi.v6i2.1348](https://doi.org/10.34306/abdi.v6i2.1348).
- [2] I. P. Gustiah, N. Lutfiani, M. R. Kusuma, E. P. Lestari, and T. Green, "Grant Submission Monitoring System Based on Orange Technology Using Laravel 12 and Vue.js," *J. Orange Technol.*, vol. 1, no. 2, pp. 75–90, Apr. 2025, doi: [10.68012/jot.v1i2.24](https://doi.org/10.68012/jot.v1i2.24).
- [3] S. D. Sugiyanti, R. Widayanti, M. B. Ulum, G. Firmansyah, and A. H. Azizah, "Design Dashboard Monitoring Teacher Performance Assessment at Cinta Kasih Tzu Chi High School," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 46–56, Oct. 2022, doi: [10.34306/itsdi.v4i1.569](https://doi.org/10.34306/itsdi.v4i1.569).
- [4] S. Suriyanto, A. Purkon, W. Aswawarman, T. S. Soeparyanto, and A. Kadir, "Performance Evaluation of Costs and Schedules for the Merah Putih Cooperative Construction Project Using Earned Value Management," *ADI Bisnis Digit. Interdisiplin J.*, vol. 7, no. 1, pp. 105–116, Jun. 2026, doi: [10.34306/abdi.v7i1.1483](https://doi.org/10.34306/abdi.v7i1.1483).
- [5] N. P. L. Santoso, R. A. Sunarjo, and I. S. Fadli, "Analyzing the Factors Influencing the Success of Business Incubation Programs: A SmartPLS Approach," *ADI J. Recent Innov.*, vol. 5, no. 1, pp. 60–71, Sep. 2023, doi: [10.34306/ajri.v5i1.985](https://doi.org/10.34306/ajri.v5i1.985).
- [6] L. Larisang, S. Sanusi, M. A. Bora, and A. Hamid, "Practicality and Effectiveness of New Technopreneurship Incubator Model in The Digitalization Era," *Aptisi Trans. Technopreneurship*, vol. 7, no. 2, pp. 318–333, Jul. 2025, doi: [10.34306/att.v7i2.482](https://doi.org/10.34306/att.v7i2.482).

- [7] A. Nuche, O. Sy, and J. Carlos Rodriguez, "Optimizing Efficiency Through Sustainable Strategies: The Role of Management and Monitoring in Achieving Goals," *Aptisi Trans. Manage.*, vol. 8, no. 2, pp. 167–174, May 2024, doi: [10.33050/atm.v8i2.2257](https://doi.org/10.33050/atm.v8i2.2257).
- [8] Sudaryono, U. Rahardja, and N. Lutfiani, "The Strategy of Improving Project Management Using Indicator Measurement Factor Analysis (IMF) Method," *J. Phys.: Conf. Ser.*, vol. 1477, no. 3, Art. no. 032023, Mar. 2020, doi: [10.1088/1742-6596/1477/3/032023](https://doi.org/10.1088/1742-6596/1477/3/032023).
- [9] N. Lachlan, R. F. Terizla, and A. Lansonja, "Rethinking IMF: An Evolutionary Concept Analysis and Redefinition of Indicator Measurement Factor Analysis for Creative-Campus Projects," *CreatiCampus Working Paper Series*, vol. 1, no. 1, Esquine CreatiResearch Institute, Oct. 2026. [Online]. Available: <https://creaticampus.esquine.com/papers/rethinking-imf-indicator-measurement-factor-analysis-creative-campus/>
- [10] E. A. Beldiq, G. Jacqueline, and A. Husain, "IMF as the Operating System of CreatiCampus Projects: A Process Framework from POAC to the Viewboard and the Creative-Campus Project Lifecycle," *CreatiCampus Working Paper Series*, vol. 1, no. 1, Esquine CreatiResearch Institute, Oct. 2026. [Online]. Available: <https://creaticampus.esquine.com/papers/imf-operating-system-creative-campus-projects-poac-viewboard-lifecycle/>
- [11] N. Ani, A. Erica, and B. Callula, "Measuring What IMF Manages: A Design-Science Protocol for an Indicator Set and Maturity Model for CreatiCampus Projects," *CreatiCampus Working Paper Series*, vol. 1, no. 1, Esquine CreatiResearch Institute, Oct. 2026. [Online]. Available: <https://creaticampus.esquine.com/papers/measuring-what-imf-manages-indicator-set-maturity-model-creative-campus/>
- [12] N. Lutfiani, U. Rahardja, and K. T. Khasanah, "The Development Viewboard As an Information Media at Official Site Asosiation," *Aptisi Trans. Manage.*, vol. 6, no. 1, pp. 10–18, Jan. 2022, doi: [10.33050/atm.v6i1.1529](https://doi.org/10.33050/atm.v6i1.1529).
- [13] D. Abbas, K. Siahaan, and M. Yusup, "Design Thinking as a Business Model for Empowering Creative Entrepreneurs in the Digital Era," *Startupreneur Bus. Digit.*, vol. 4, no. 2, pp. 124–133, Oct. 2025, doi: [10.33050/sabda.v4i2.805](https://doi.org/10.33050/sabda.v4i2.805).
- [14] F. Zidan, D. Nugroho, R. Asri, and S. Agustina, "Peran Pendidikan dalam Mempersiapkan Generasi Creativepreneur di Masa Depan," *ADI Bisnis Digit. Interdisiplin J.*, vol. 5, no. 1, pp. 41–46, Jun. 2024, doi: [10.34306/abdi.v5i1.1099](https://doi.org/10.34306/abdi.v5i1.1099).
- [15] F. Rahardja, L. K. Choi, R. C. Wijaya, and R. A. Sunarjo, "Gamification in Digital Startups: Enhancing User Engagement and Business Growth," *Startupreneur Bus. Digit.*, vol. 4, no. 1, pp. 1–11, Apr. 2025, doi: [10.33050/sabda.v4i1.639](https://doi.org/10.33050/sabda.v4i1.639).
- [16] A. S. Bist, B. Rawat, U. Rahardja, Q. Aini, and A. G. Prawiyogi, "An Exhaustive Analysis of Stress on Faculty Members Engaged in Higher Education," *IAIC Trans. Sustain. Digit. Innov.*, vol. 3, no. 2, pp. 126–135, Apr. 2022, doi: [10.34306/itsdi.v3i2.533](https://doi.org/10.34306/itsdi.v3i2.533).
- [17] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012).
- [18] M. H. R. Chakim, U. Rahardja, E. D. Astuti, E. Erika, and C. T. Hua, "The Social Empowerment Role of the Penta Helix Entrepreneurship Ecosystem in Driving the National Economy," *ADI Pengabd. Masy. J.*, vol. 6, no. 1, pp. 1–13, Sep. 2025, doi: [10.34306/adimas.v6i1.1283](https://doi.org/10.34306/adimas.v6i1.1283).
- [19] R. Ginting Munthe, Q. Aini, N. Lutfiani, I. Van Persie, and A. Ramadhan, "Transforming Scientific Publication Management in the Era of Disruption: SmartPLS Approach in Innovation and Efficiency Analysis," *Aptisi Trans. Manage.*, vol. 8, no. 2, pp. 123–130, May 2024, doi: [10.33050/atm.v8i2.2261](https://doi.org/10.33050/atm.v8i2.2261).
- [20] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979, doi: [10.1037/0033-2909.86.2.420](https://doi.org/10.1037/0033-2909.86.2.420).
- [21] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychol. Methods*, vol. 1, no. 1, pp. 30–46, Mar. 1996, doi: [10.1037/1082-989X.1.1.30](https://doi.org/10.1037/1082-989X.1.1.30).

- [22] A. F. Hayes and K. Krippendorff, “Answering the Call for a Standard Reliability Measure for Coding Data,” *Commun. Methods Meas.*, vol. 1, no. 1, pp. 77–89, Apr. 2007, doi: [10.1080/19312450709336664](https://doi.org/10.1080/19312450709336664).
- [23] J. M. Bland and D. G. Altman, “Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement,” *Lancet*, vol. 327, no. 8476, pp. 307–310, Feb. 1986, doi: [10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8).
- [24] S. Wibowo, I. A. Widjaya, J. Zanubiya, R. Evans, and U. Rahardja, “Orange Technology for Humanistic Innovation in Higher Education,” *J. MENTARI*, vol. 4, no. 2, pp. 105–115, Mar. 2026, doi: [10.33050/mentari.v4i2.899](https://doi.org/10.33050/mentari.v4i2.899).
- [25] U. Rahardja, N. Ani, and N. Leonie, “‘Inclusively Creative’ Campuses: An Orange Technology Framework for Humanistic Creative-Economy Higher Education in Indonesia,” *CreatiCampus Working Paper Series*, vol. 1, no. 1, Esquine CreatiResearch Institute, Oct. 2026. [Online]. Available: <https://creaticampus.esquine.com/papers/inclusively-creative-campus-orange-technology-wcce-2026/>
- [26] M. Hardini, M. Mardiana, D. Andayani, E. D. Astuti, and I. Maria, “Evaluation of Digital Branding Strategy in OJS-Based Journals to Increase Dissemination of Scientific Publications,” *ADI J. Recent Innov.*, vol. 8, no. 1, pp. 1–12, Sep. 2026, doi: [10.34306/ajri.v8i1.1495](https://doi.org/10.34306/ajri.v8i1.1495).
- [27] E. P. Lestari, A. Valerry, and L. Sulistyowati, “Measuring the Effect of Interdisciplinary Learning Factory Projects on Student Learning Outcomes,” *Int. Trans. Educ. Technol.*, vol. 4, no. 2, pp. 131–148, May 2026, doi: [10.33050/itee.v4i2.1109](https://doi.org/10.33050/itee.v4i2.1109).
- [28] I. K. Mertayasa, U. Rahardja, M. Hardini, and E. P. Harahap, “Interior Design Alphabet Incubator 3.0 Based on Planner 5D,” *Technomedia J.*, vol. 8, no. 1, pp. 82–108, May 2023, doi: [10.33050/tmj.v8i1.1982](https://doi.org/10.33050/tmj.v8i1.1982).