

---

# Do University Rankings Show Which Universities Teach Better?

## *A Claim-Level Framework for Evaluating Educational Interpretations*

Thanakit Ouanhlee

Thipsamai Research Institute, Thailand

ORCID: <https://orcid.org/0000-0002-8449-4145>

### Abstract

University rankings combine research performance with evidence about resources, student experience and educational outcomes, and their results can be read as statements about teaching. This conceptual article asks whether such results can show which universities teach better. Drawing on argument-based validation, it proposes a claim record that links each educational statement to its evidence and applies three checks: whether the evidence addresses the stated construct, whether its scope supports the population and any comparison implied, and whether the design supports the strength of the inference. Each statement is retained or withheld, and any narrower version must pass all three checks before reporting. The procedure is applied to seven documented indicators from the Times Higher Education (THE) and QS rankings, to three stipulated cases and, as illustrations, to two documented public statements: one describing a Teaching Excellence Framework Gold rating and one describing a ranking's teaching-pillar score. Research indicators, resources and reputation support useful descriptions but do not, without additional evidence, establish learning outcomes or institutional contribution. The ranking indicators examined therefore cannot show that one institution produces more undergraduate learning than another. Keeping the meaning and evidence of research and educational results separately inspectable is necessary but not sufficient; educational comparison also requires wording bounded to the evidence, evidence at the level of the claim made, and explicit uncertainty. Users may still weight results by their own priorities, provided the resulting decision score is not presented as demonstrated teaching effectiveness. The framework's practical effects remain to be tested empirically.

Keywords: university rankings; teaching effectiveness; research assessment; learning gains; argument-based validation; composite indicators; Teaching Excellence Framework

## 1 Introduction

A university ranking can be arithmetically correct while supporting a narrower conclusion than the phrase "best university" suggests. A composite score orders institutions according to selected indicators and weights. The resulting order does not establish that the highest-ranked institution provides the strongest education for a particular student, programme or level of study. An assessment must therefore be judged partly by the relationship between its evidence and the conclusions that users are invited to draw.

The distinction matters because research performance and educational effectiveness answer different questions. Research assessment concerns contributions to knowledge and their influence. Educational assessment concerns the organisation of learning and what students experience, achieve and develop. The activities can interact, but a connection between activities does not make their indicators interchangeable. A publication count cannot, by itself, establish what undergraduates have learned.

The problem has substantial precedents. Ioannidis et al. (2007) examined the validity of criteria used to represent educational and research excellence in international rankings. In an exploratory analysis of data from the Academic Ranking of World Universities (ARWU), Times Higher Education (THE) and QS Quacquarelli Symonds (QS) rankings for 2012–2018, Selten et al. (2020) identified reputation and research performance as major underlying dimensions, alongside other dimensions and variation across systems. Their analysis concerns the structure of ranking indicators rather than observations of student learning, and it does not establish that every system or later edition shares that structure, but it supports asking what an assessment represents.

Multidimensional ranking design and principles for the responsible use of rankings also establish the separation of performance dimensions, as discussed in Section 3. An argument for separation must build on these contributions rather than claim that the principle is new.

The central question is how an assessment can report educational claims without treating research performance, resources or reputation as sufficient evidence of teaching effectiveness. The article answers by proposing a claim record that connects each educational statement to its evidence, the assumptions required to interpret it and the wording that survives examination. The contribution is an adaptation of established validity theory rather than a new theory: it moves the unit of scrutiny from the indicator or the ranking system to the individual statement a reader is invited to accept, and it specifies what that statement must satisfy before it can be retained. Applied consistently, the procedure also yields a substantive conclusion. Unless combined with evidence on student learning, the ranking indicators examined here cannot support a claim that one institution produces more undergraduate learning than another, whereas narrower descriptive statements, made at the level the evidence covers, often can. The article examines the risk of unwarranted inference from indicators to educational conclusions; it does not estimate how often providers or users draw such inferences.

### 1.1 Approach and scope

The article develops a conceptual framework informed by argument-based validation and applies it to the interpretation of university comparisons. It is a conceptual article based on purposively selected literature, not a systematic review. Sources were located through public web searches in October 2026 on rankings, teaching effectiveness, the research–teaching relationship, learning gain, value added, composite weighting and ranking interpretation, and were obtained from publishers, repositories, author manuscripts, preprints and official documentation. They were selected for

conceptual relevance, including evidence that challenges separation. Where a preprint, working paper or open version was consulted, the reference identifies that version and page numbers refer to it; where only an abstract was available, as for Kane (2013), only claims stated in the abstract are used. No comprehensive database search or formal screening was undertaken, so exhaustive coverage cannot be claimed.

Empirical studies inform the distinctions between research, experience and learning; validity theory and ranking critiques inform the reporting proposal. The reasoning identifies the claim an indicator could support, exposes the additional assumptions needed for a stronger claim and examines whether revised wording resolves the mismatch. Official documents define the seven indicators examined in Section 5.2, which are purposively selected illustrations rather than a representative sample. Three stipulated cases in Section 5.4 make the decision rules inspectable, and Section 5.5 illustrates their application to two documented public statements. None of these examples measures institutional effectiveness or user behaviour.

The educational focus is undergraduate provision, with programme or subject comparisons preferred where the evidence permits. Postgraduate and doctoral education involve different research–teaching relationships and require separate analysis. Engagement, professional service and cultural contribution may also warrant distinct assessment; the two-domain focus isolates the question under discussion without defining the whole university through two scores.

## 2 Distinguishing educational claims from research performance

Educational quality is too broad to be represented adequately by an unspecified teaching score. Resources, instructional practices, student experience, attainment and progress describe related but different features. A well-equipped department may provide conditions favourable to learning. A student survey may describe perceived clarity or support. Neither observation alone demonstrates that students acquired more knowledge than they would have acquired elsewhere.

The framework adopts teaching effectiveness as a claim about how educational provision contributes to intended learning. Establishing that contribution is more demanding than describing provision or recording final attainment. Not every useful educational indicator must establish causality; however, its label should identify what it actually measures. Table 1 distinguishes the principal claims needed for this purpose.

**Table 1 Educational dimensions and the questions they answer**

| Dimension                  | Question addressed                                                        |
|----------------------------|---------------------------------------------------------------------------|
| Resources and conditions   | What staff, facilities and support are available?                         |
| Instructional practice     | How are curricula, activities, feedback and assessment organised?         |
| Student experience         | How do participating students describe their experience?                  |
| Attainment                 | What can students demonstrate at a specified point?                       |
| Observed learning gain     | What change is measured over a specified period?                          |
| Institutional contribution | What part of the outcome can defensibly be attributed to the institution? |

Source: proposed analytical distinctions, informed by Gibbs (2010), Rogaten et al. (2019), Vermunt et al. (2018) and Cifuentes Gomez and Santelices (2024).

The first row corresponds broadly to what Gibbs (2010, pp. 4, 12), adapting Biggs's 3P model, calls presage variables, and the second and third to process variables, since experience surveys report on what happens while students learn; the remaining rows divide his product category into attainment, observed gain and institutional contribution. The division matters because Gibbs's synthesis concluded that presage variables such as funding, research performance and the reputation that enables selective entry explain little of the variation between institutions in educational gains, while product measures such as grades largely reflect the prior attainment of the students those institutions attract (p. 5). The review is principally concerned with the United Kingdom and relies heavily on studies from the United States; it is used here for its classification and the direction of its findings, not as a current estimate for any ranking.

The distinctions are consistent with research on learning-gain measurement. Rogaten et al. (2019) review affective, behavioural and cognitive measures. They caution against comparisons across heterogeneous designs and found no randomised controlled trials, so their summaries do not establish causal institutional effects. Vermunt et al. (2018) develop a conceptual framework and instrument using the first wave of a planned longitudinal survey in research-intensive English universities, analysing complete survey responses. That paper does not establish gains over time, and some scales require further development. Wider institutional comparisons would require further evidence of coverage and comparability. Constructing and validating measures is itself a substantive research task, and validating a scale does not validate an institutional league table built from it.

Final attainment is especially vulnerable to an interpretation that confuses selection with contribution. A university enrolling highly prepared students may produce high final scores without producing the greatest progress, while one serving less prepared students may contribute substantially despite lower final attainment. These are analytical possibilities, not claims about particular institutions.

Cifuentes Gomez and Santelices (2024) show that estimates of institutional contribution depend on outcome definitions and the treatment of selectivity. Their review also cautions that much of the relevant evidence remains descriptive. Statistical adjustment can improve a comparison, but an adjusted difference should not be called a causal teaching effect unless the research design and assumptions justify that interpretation.

Student feedback should be retained for the questions it can answer. Uttl et al. (2017) challenge the use of student teaching ratings as evidence of learning: their meta-analysis reports positive pooled correlations that become substantially weaker when prior learning and small-study effects are addressed. The included studies correlated section-level ratings with common assessments within courses, so they do not directly estimate differences between whole universities. The finding warrants caution about substituting ratings for learning measures, while reports of accessibility, feedback and learning conditions remain useful in their own right, subject to checks of response coverage and possible bias.

## **2.1 Research and teaching can interact without sharing a measure**

Separating assessment does not require universities to stop conducting research, nor does it require researchers and teachers to work in separate institutions. The concern is the evidential meaning of the assessment. Research activity may inform curricula, expose students to developing knowledge and support inquiry, while its educational benefits remain questions for investigation.

The meaning of teaching effectiveness also needs a boundary. Here, educational provision includes curriculum, assessment and learning support as well as classroom instruction. An estimated effect of attending an institution may include peer interactions and other features of its environment. Without a design that distinguishes their contribution, such an estimate cannot be attributed to individual instructors. The institution, programme and teacher are therefore different objects of assessment, even when the same outcome is of interest.

Hattie and Marsh (1996) synthesised 58 studies containing 498 correlations and reported a weighted mean correlation of approximately .06 between measures of research and teaching; in a later account of their research programme, the authors described this as less than 1% of variance in common (Hattie and Marsh, 2004, p. 2). The finding challenges the assumption that research productivity guarantees effective teaching, but the authors have been explicit about its limits. Their analyses concerned individual academics and departments rather than whole institutions, and they rejected the use of the finding to justify separating research and teaching for funding purposes, treating any such separation as a matter of institutional mission rather than an implication of a zero correlation (Hattie and Marsh, 2004, pp. 1, 6–7). They also proposed ways of strengthening the research–teaching relationship within institutions (Hattie and Marsh, 2004, p. 7). The level of analysis matters for the present argument: a correlation among academics cannot by itself support a claim about which university teaches better.

Palali et al. (2018), whose analysis is also available as an open discussion paper (Palali et al., 2017), used random allocation to tutorial groups at Maastricht University's School of Business and Economics. In baseline models, master's students assigned to teachers with higher-quality publication records obtained higher course grades; bachelor's students did not. Student evaluations did not follow the pattern of grades. Random assignment reduces student sorting, but it does not randomise teachers' publication records independently of their other characteristics. The analytical sample was also restricted to evaluation respondents and teachers with available publication records (Palali et al., 2017, pp. 10–11). Adding academic-position controls removed statistical significance, which the authors attribute to the high correlation between publication records and position and the resulting loss of precision (Palali et al., 2017, p. 16); this sensitivity analysis does not, by itself, show no effect. The results therefore concern assignment to teachers with particular research records, not the effect of publishing more. The authors state that their findings do not determine whether research and teaching should be organisationally separated (Palali et al., 2017, p. 18).

Healey (2005) argues for active student involvement in inquiry as a way to connect research and learning, offering a conceptual synthesis rather than a causal estimate. Gibbs (2010, pp. 28–29), summarising earlier studies, reaches a compatible conclusion: measures of a department's research have been found not to correlate with measures of its teaching, and where research environments benefit undergraduates, the benefit is deliberately engineered, so that the relevant indicator is an undergraduate research programme rather than research strength. Both accounts locate the educational value of research in pedagogy, not publication volume. Under the proposed framework, an evaluated inquiry activity contributes to educational evidence and its publications to the research profile; separate reporting is this article's proposal, not a recommendation attributed to these studies.

## 2.2 Why a relationship does not make indicators interchangeable

The argument for separate assessment does not depend on research and teaching being unrelated. Even a strong positive association would leave open what a research indicator predicts in the setting assessed and what the resulting estimate permits a reader to conclude. Evidence that research

productivity predicts an educational outcome could justify its use within a validated predictive model, reported with its population, predictive error and uncertainty; it would not make publication volume a direct observation of learning or show that more publications would improve teaching, and a relationship found among individual teachers cannot simply be transferred to university-wide comparisons. Prediction, observed outcome and institutional contribution should therefore remain distinguishable in the report. The argument rejects unexplained substitution rather than all indirect evidence, and separate reporting makes the inferential step visible and open to challenge.

### 2.3 Interpretation and use as the objects of validation

Argument-based validation supplies a direct foundation. Kane (1990, pp. 1–2, 27, 34–37) connects evidence requirements to the assumptions linking scores with interpretations and actions, allows interpretations to be revised or abandoned, and treats validation judgements as provisional. The Standards for Educational and Psychological Testing define validity as the degree to which evidence and theory support interpretations of scores for proposed uses, require each intended interpretation to be validated, and ask that users be cautioned when a common or likely interpretation has not been evaluated or conflicts with the available evidence (American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 2014, pp. 11, 23). The abstract of Kane (2013) likewise holds that stronger claims require stronger support. The focus on individual interpretations, and the possibility of revising them, are therefore inherited rather than introduced here; the Standards' requirement to caution users is close to what the present procedure asks of ranking providers.

The proposed adaptation concerns reporting across university assessment domains. Construct asks whether the observation addresses the stated educational attribute. Scope and comparison ask how far the evidence extends across participants, programmes, periods and institutions. Inference asks whether the conclusion describes, predicts or attributes an outcome, and whether the evidence supports that step. These are practical prompts, not independent types of validity or an exhaustive validation theory.

The three checks broadly correspond to the kinds of inference that Kane (1990, pp. 15–17) identifies in interpretive arguments. The construct check concerns extrapolation from the observations to the attribute named in a claim, together with the theory-based assumptions that link them; the scope-and-comparison check concerns generalisation across occasions, participants and settings, and extrapolation beyond those observed; and the inference check concerns whether a claim adds a prediction, a causal explanation or a decision, each resting on further assumptions. Because Kane notes that generalisation and extrapolation are not sharply distinct (p. 16), a threat may be recorded under each check it affects, without repeated entries being treated as additional evidence. His observation that a questionable assumption can sometimes be dropped by limiting the interpretation (p. 27) is the operation formalised here as NARROW.

An acceptable description also does not settle whether using it for a particular decision is justified. Admissions, funding and institutional choice involve purposes, consequences and values beyond the meaning of a score. The record proposed here makes educational wording inspectable; the responsibilities discussed in Section 6.3 address its subsequent use. A completed record therefore cannot certify every decision based on the information.

### 3 Existing responses and the remaining design problem

Existing responses already provide routes beyond an undifferentiated ranking: multidimensional assessment, education-focused assessment and responsible-ranking principles. Together they establish a substantial foundation for the present proposal.

The CHE Centre for Higher Education (CHE) description of the 2022 U-Multirank edition identifies five dimensions: teaching and learning, research, knowledge transfer, international orientation, and regional engagement. It describes performance grades for different activities and comparisons guided by users' interests (Federkeil, 2022). The U-Multirank design therefore provides a documented precedent for multidimensional reporting, not evidence that its educational indicators measure causal learning effects. The Teaching Excellence Framework (TEF) in England provides a different comparator: its 2023 framework provides an overall rating alongside separate ratings for student experience and student outcomes (Office for Students, 2023). TEF ratings are not a worldwide ordinal league table, and they should not be presented as a direct causal estimate of institutional educational contribution.

Alternative weighting is a competing response. De Witte and Hudrlikova (2013) propose a benevolent composite ranking in which institution-specific weights favour relative strengths, and they treat the absence of a single result in indicator-based approaches such as Leiden and U-Multirank as a disadvantage. Their argument is therefore a counterposition to separate reporting. The present proposal accepts that users may want a combined decision score but requires its preference-dependent meaning to remain visible; and because their application retains the staff-to-student ratio as a teaching proxy, reweighting it does not supply evidence of learning.

Waltman et al. (2017) reject a generic concept of university performance and the compensation of weakness in one dimension by strength in another, and Gadd et al. (2021) include indicator validity, institutional mission, avoiding composite indicators and tailoring information to audiences among the criteria they apply to six ranking systems. These are close precedents: the present framework is an educational application of established principles, not the invention of domain separation or of caution about aggregation.

Vernon et al. (2018) review research-related ranking systems and identify limitations in measures of academic quality, although their eligibility criteria required research measures and so cannot show that all rankings neglect education. Srisawat (2020) discusses the research emphasis of major international rankings and Thai universities' strategic responses; this local precedent limits claims of novelty about the concern without endorsing the present solution.

The remaining design problem occurs at the level of the claim. A platform can separate research and education into different columns while using an educational label that exceeds what its indicators establish. Resources remain resources when placed under a teaching heading. Conversely, a carefully bounded account of resources may be useful even when no learning outcome is measured. Domain separation addresses what is combined; examination of individual claims addresses what the components mean. Both are required for an interpretable assessment.

The proposal therefore sits at the intersection of three precedents. Multidimensional reporting separates the information presented; responsible-ranking guidance evaluates assessment systems and their indicators; argument-based validation examines the justification of interpretations. The present procedure brings these concerns into a single record containing one educational claim, three reasoned checks, a disposition and any revised wording.

### 3.1 What the reporting procedure adds

The procedure adds three operations to these precedents. First, it makes the individual statement the unit of assessment: a staffing ratio can remain in an education profile while a claim about superior learning drawn from it is withheld. Second, it requires an explicit disposition for each statement, with any narrower wording recorded as a linked revision and reassessed rather than silently substituted. Third, it states minimum evidential requirements by type of claim – descriptive, comparative, predictive and causal – so that sufficiency is argued rather than asserted. The underlying validity principles are inherited; the contribution is their specification for the boundary between research performance and educational conclusions, together with the conclusions about comparative teaching claims that follow from applying it (Section 5.6).

An exit-score example illustrates the operation. A profile labelled "education" could report a high exit score without clarifying whether it represents attainment or institutional contribution. A completed record would retain the attainment description while identifying the additional assumptions required for attribution. Where existing guidance already reaches that verdict, the record supplies a documentation format rather than a different judgement; whether it improves consistency or comprehension is an empirical question.

## 4 A framework for separate reporting and bounded claims

The proposed framework begins with the intended question, then identifies the evidence needed to answer it. A prospective undergraduate may ask about learning and support in a particular programme. A research funder may ask about scholarly contributions in a field. A university governing body may require both profiles. Shared users and institutional boundaries do not require those questions to be answered by a shared score.

The research profile should identify the research dimensions being assessed, such as outputs, scholarly influence or assessed contributions, with their coverage and limitations. The education profile should distinguish the dimensions in Table 1. Neither profile should be labelled comprehensive unless its content justifies that description. A domain can contain several measures, and separate reporting need not result in two new composite league tables.

### 4.1 Match each claim to its evidence

For each proposed educational claim, record the evidence source and methodology version, measurement method, population, unit, period and intended comparison. Include coverage, missingness, validity evidence, adjustments and uncertainty, with a page, table or section locator. Table 2 summarises the proposed relationships between evidence and interpretation. The claim is the unit of assessment: one indicator can support a limited description while failing to support a broader conclusion.

**Table 2: Evidence and the limits of educational claims**

| Evidence                   | Potentially supported interpretation                 | Stronger claim requiring more evidence          |
|----------------------------|------------------------------------------------------|-------------------------------------------------|
| Publications and citations | Research output or influence within stated coverage. | Better instruction or greater student learning. |

| <b>Evidence</b>                      | <b>Potentially supported interpretation</b>               | <b>Stronger claim requiring more evidence</b>                           |
|--------------------------------------|-----------------------------------------------------------|-------------------------------------------------------------------------|
| Staffing and resources               | Conditions potentially supporting education.              | Resources have improved learning.                                       |
| Experience surveys                   | Reported experience among respondents.                    | Direct knowledge gain or a representative view without coverage checks. |
| Moderated student work               | Attainment on assessed tasks.                             | The university caused the attainment.                                   |
| Baseline and follow-up assessments   | Observed progress on comparable measures.                 | Greater causal contribution than another institution.                   |
| Evaluation of inquiry-based learning | Educational findings within the study's design and scope. | Benefits from research output alone or across all programmes.           |

The middle column states interpretations that may be supported when measurement, coverage and data quality are adequate; it does not certify evidence by its label, and the rows are not a hierarchy in which causal estimates are always preferable. The relevant test is whether the evidence answers the question actually posed.

The assessor applies the three checks to each proposed statement: whether the measure addresses the stated construct, whether the population, unit, period and measurement conditions support its scope and any comparison, and whether the design justifies the strength of the inference. A descriptive statement requires descriptive support; a causal statement requires a defensible basis for attributing an outcome to the institution. A statement about one cohort needs no cross-institutional comparison, but its population and period must still be specified.

For each check, record sufficient, insufficient or indeterminate support, together with the reason and evidence locator. "Insufficient" means that the documented evidence does not justify the statement, for example because its design cannot identify the claimed causal effect. "Indeterminate" means that necessary information is unavailable, so the relevant support cannot be established. Neither status authorises the statement. Retain the original wording only when all applicable checks are sufficient. Otherwise withhold that wording. Where a restricted claim is potentially defensible, record a linked revision and apply all three checks again. NARROW is this revision step, not a final judgement that the revised claim is valid. Figure 1 shows the procedure.

Sufficiency is judged against requirements stated for the particular claim rather than a universal numerical threshold. A record identifies the intended attribute, the observations used to represent it, the target population and period, and the assumptions connecting evidence to wording; it also explains which material threats have been addressed and which unresolved threats would change the conclusion. Where no comparison is made, the comparison component is marked not applicable with a reason; a checklist entry without such justification is not sufficient support.

The requirements vary with the type of statement and are summarised in Table 3, which sets out, for each type of claim, the minimum record and the conclusions that the claim type does not support by itself. Comparative, predictive and causal claims each add requirements specific to the inference they make, and none of the stronger types is established merely by satisfying the requirements of a

weaker one. Baseline measurement may assist causal identification but is not a universal requirement.

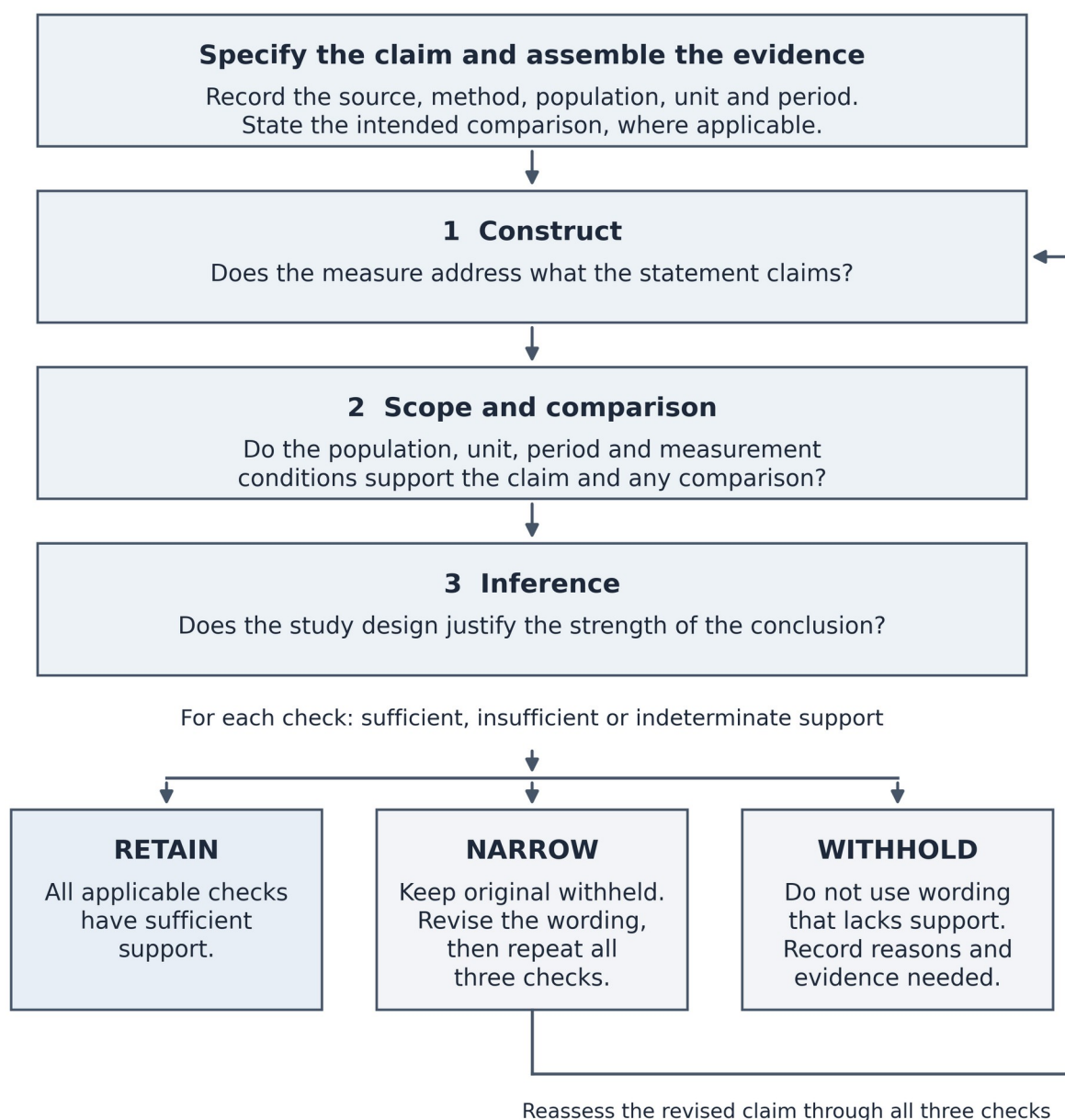
**Table 3 Minimum evidential requirements by type of claim**

| Type of claim | Minimum record                                                                                                                                                                  | Not established by this claim type alone                                           |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Descriptive   | Measure or question; administration or scoring; observed group and denominator; period; relevant missingness.                                                                   | Representativeness beyond the observed group; comparison; prediction; attribution. |
| Comparative   | Descriptive requirements plus a common or defensibly linked construct and scale, compatible conditions and populations, and an account of precision and contextual differences. | Why the difference arises; institutional contribution.                             |
| Predictive    | Validation for the intended population and use, including predictive error.                                                                                                     | That the predictor directly observes the outcome; institutional contribution.      |
| Causal        | An explicit counterfactual and a design-based argument addressing the main competing explanations.                                                                              | Generalisation beyond the population and conditions of the design.                 |

Source: proposed reporting requirements summarising Section 4.1; they are not validated cut-offs.

Assessors should specify material threats, and the evidence needed to address them, before examining which institution appears to perform better. Missing disclosure makes the affected check indeterminate; a design demonstrably incapable of the proposed inference makes it insufficient, however large the dataset. A disputed criterion should be referred to a suitably qualified reviewer, with the competing reasons preserved and the wording withheld if the dispute changes whether the claim is supportable. Agreement among reviewers cannot replace an evidential justification.

Each record contains the claim identifier and exact wording; whether the statement is quoted from a source or constructed for analysis; the evidence locator and version; the construct, population, unit and period; the result and reason for each check; the final wording; and unresolved questions. A revision receives a linked identifier so that a change in meaning cannot be hidden as stylistic editing. Narrowing should preserve a recognisable part of the original informational purpose. Replacing a claim about teaching effectiveness with an unrelated research claim is a change of domain, not a successful educational revision.

**Figure 1 Assessing educational claims against their evidence**

Source: author-proposed reporting procedure, informed by argument-based validation and responsible-ranking principles. Unsupported original wording remains withheld during revision. NARROW is an intermediate step; revised wording must pass all three checks before use. Record the reasons, uncertainty and evidence needed for reconsideration.

The completed records in Section 5.4 illustrate these requirements under stipulated conditions: a descriptive learning-gain claim that can be retained, an overbroad survey claim that can be narrowed and an educational comparison that must be withheld. Section 5.5 illustrates the same requirements with two documented public statements.

#### 4.2 Keep the unit of evidence aligned with the unit of comparison

A whole-university research result cannot establish the quality of a particular programme, and evidence from one course or a small participant group cannot establish effectiveness across an institution. Reports should identify the level to which the evidence applies and explain any aggregation across courses, programmes or cohorts.

Comparison groups also need justification. Disciplines assess different capabilities, and degree levels have different aims, so comparability may be stronger within a subject and qualification level than across all universities. Broad comparison remains possible, but the assessor must explain what remains comparable and what information is lost.

Institutional context should inform interpretation without becoming an automatic excuse for weak outcomes. Prior attainment, student composition and programme characteristics can help explain differences, provided the chosen adjustments and their effects are visible and unadjusted descriptions are kept distinct from contextualised estimates.

### **4.3 Prevent compensation across domains**

The central reporting recommendation is that research achievement should not earn credit in a score presented as teaching effectiveness merely because it is valued as research. Where research contributes to learning, the education profile should record that contribution through educational evidence; publication and citation counts remain research indicators. A validated educational prediction of the kind discussed in Section 2.2 is a different claim, and should be labelled as an estimate rather than an institutional causal contribution. Recording each contribution only in the domain whose evidence supports it preserves the meaning of each assessment while allowing the same academic activity to serve both missions.

An education profile should report any unresolved evidence gap alongside the research result (Section 4.4); neither domain supplies missing evidence for the other.

Users may legitimately value both domains when choosing an institution. A separate decision-support tool could let them state those preferences, provided any combined value is labelled as a preference-dependent decision score with visible weights, not as a teaching-quality measure or universal institutional merit, and the underlying domain results remain accessible.

### **4.4 Report missing evidence and uncertainty**

Missing educational information should be identified as missing. Substituting reputation or research output without a validated relationship to the educational outcome would change the question being answered. If a model is used to estimate an unobserved outcome, both the missing observation and the estimate should remain visible. Assigning a zero would also be misleading if it implied demonstrated poor performance. Where participation differs across institutions, the report should show the observed coverage and explain how that affects comparison.

Uncertainty includes sampling variation, measurement limitations, selection effects and sensitivity to analytical choices. A confidence interval may address some of these sources without resolving the others. Reports should explain which uncertainty has been quantified and which remains qualitative. Bands can sometimes communicate limited precision more appropriately than exact ranks, but their thresholds also require justification.

The framework therefore treats "insufficient evidence for comparison" as a legitimate result. Declining to order two institutions can be more informative than producing an exact-looking difference that the evidence cannot sustain.

Treating missing evidence cautiously must not make non-disclosure advantageous. Under the proposed procedure, providers should distinguish outcomes that were not measured from data that were collected but not supplied, or supplied without adequate documentation. Where a reporting requirement applies, compliance can be assessed separately. Failure to meet that requirement may

justify a finding about transparency or reporting practice; it does not establish the unobserved level of learning. Recording the reason for the gap preserves accountability without inventing an educational result.

#### 4.5 Interpret multiple sources without averaging away disagreement

An education profile may contain evidence pointing in different directions. Strong attainment alongside dissatisfied students is not necessarily a contradiction, because capability and experience are different objects of assessment, and both findings should be preserved. Where two sources address the same outcome but disagree, their cohorts, periods, measurement conditions and coverage should be examined before they are combined, since averaging incompatible observations can hide the very difference that requires explanation.

A hypothetical programme might show rising test scores among continuing students while losing many participants before the final assessment. The increase supports a statement about the observed group if the measurement is sound, but not about the entering cohort; completion records and information about non-participants qualify the scope of the claim. Additional evidence improves interpretation by addressing a specific uncertainty, not by adding to a count of favourable indicators.

The same reasoning limits composite educational scores. A profile can report disciplinary knowledge, practical capability and student experience without declaring them interchangeable; combining them requires an account of the trade-offs and of outcomes for which compensation is unacceptable. The framework makes those choices explicit without supplying a universal definition of a good education.

### 5 Applications of the reporting framework

#### 5.1 A hypothetical aggregation example

Consider two hypothetical institutions with research and educational scores already expressed on comparable 0–100 scales. University A has scores of 90 and 50 respectively; University B has scores of 50 and 80. These invented values simplify the arithmetic. They do not establish that research or education can, in practice, be measured with equivalent precision. Table 4 sets out these scores and the combined result under a 70:30 weighting.

**Table 4 Hypothetical institutional scores**

| Institution  | Research | Education | Combined |
|--------------|----------|-----------|----------|
| University A | 90       | 50        | 78       |
| University B | 50       | 80        | 59       |

Source: invented illustration. The combined score gives research a weight of 70% and education a weight of 30%.

With a research weight of 0.70 and an education weight of 0.30, A receives  $0.70 \times 90 + 0.30 \times 50 = 78$ . B receives  $0.70 \times 50 + 0.30 \times 80 = 59$ . The combined order favours A, although the educational measure favours B. The overall ranking is correct under its formula, but it does not answer which institution has the higher educational score.

Let  $w$  denote the research weight. A's combined score is  $50 + 40w$ , and B's is  $80 - 30w$ . The difference is therefore  $70w - 30$ . The institutions tie when  $w = 3/7$ , approximately 42.9%. Above

that weight A leads; below it B leads. Nothing about either institution changes when the ranking reverses. The reversal reflects a different valuation of the two domains.

The example also exposes a limit of sensitivity analysis. Showing the crossover makes the weighting choice transparent, but it does not validate either underlying score. With real indicators, a further difficulty arises: nominal weights do not reliably express the influence each component has on the final order, because that influence also depends on the indicators' variances and correlations. Paruolo et al. (2012, pp. 3–4, 16–17) demonstrate this for the 2008 Times Higher Education Supplement ranking, in which the teacher–student ratio, the component intended to capture teaching, carried a nominal weight of 20%, whereas its estimated relative importance, calculated by normalising the cross-validation main effects to sum to one, was approximately 9%. Current editions use different indicators and weights, and the analysis has not been repeated for them here, but the mechanism is general. A declared teaching weight is therefore not evidence that teaching information drives a ranking; a provider making that claim would need to report effective importance as well as nominal weights.

If A's educational score were unavailable, the research score alone would not supply the missing value. The report would show a research result and an educational evidence gap under Section 4.4, rather than infer an educational score of zero.

5.2 Interpretive boundaries of documented ranking indicators

Seven documented indicators supply illustrations: the five in the THE World University Rankings 2026 Teaching pillar, defined in the September 2025 methodology, and the QS Faculty-Student Ratio and Citations per Faculty indicators, from pages updated in November 2024 and June 2025 (Times Higher Education, 2025; QS Quacquarelli Symonds, 2024, 2025). They were selected to contrast reputation, resources, degree composition and research evidence. The exercise examines the inferential reach of documented definitions, not institutional data, provider intentions or the prevalence of misleading statements.

Both providers already differentiate domains: THE describes its Teaching pillar as the learning environment and publishes pillar scores, and QS places the staff ratio under Learning Experience and citations under Research and Discovery. The question posed here, which is not attributed to either provider, is whether the specified observations alone could justify a further claim of superior undergraduate learning. Table 5 lists each indicator with its overall weight, what it documents, and the interpretation it can and cannot support.

Table 5 Documented indicators and their interpretive boundaries

| Indicator and overall weight    | Documented observation                                             | Potential interpretation and boundary                                                                                                                     |
|---------------------------------|--------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| THE Teaching Reputation<br>15%  | Weighted academic survey mentions.                                 | Retain a claim about surveyed reputation. Withhold a claim of measured undergraduate learning.                                                            |
| THE Student Staff Ratio<br>4.5% | Full-time-equivalent (FTE) academic staff divided by FTE students. | Retain institution-level staffing density. Withhold a programme-level access claim unless allocation evidence supports it; reassess any narrower wording. |

| Indicator and overall weight       | Documented observation                                                                                    | Potential interpretation and boundary                                                                                                    |
|------------------------------------|-----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| THE Doctorate Bachelor Ratio<br>2% | Doctorates awarded relative to undergraduate degrees.                                                     | Retain degree-award composition. Withhold an inference of superior undergraduate instruction.                                            |
| THE Doctorate Staff Ratio<br>5.5%  | Subject-weighted doctoral awards relative to academic staff.                                              | Retain doctoral output relative to staffing. Withhold an inference about undergraduate educational contribution.                         |
| THE Institutional Income<br>2.5%   | Income adjusted for purchasing-power parity (PPP), relative to academic staff.                            | Retain a resource indicator. Withhold a claim that resources caused greater student learning.                                            |
| QS Faculty-Student Ratio<br>10%    | Staff-to-student ratio; includes teaching and research staff and undergraduate and postgraduate students. | Retain institution-level staffing density. Programme contact, instructional practice and learning remain unmeasured by the ratio itself. |
| QS Citations per Faculty<br>20%    | Adjusted Scopus citations relative to faculty; includes field normalisation and self-citation exclusion.  | Retain a research indicator within its coverage. Withhold an educational inference without separate evidence of student learning.        |

Sources: Times Higher Education (2025, pp. 10, 14) and QS Quacquarelli Symonds (2024, 2025). Weights refer to each system's overall ranking, not a common scale. Each row groups distinct possible claims about an indicator and is not a completed claim record. The final column gives conditional interpretations proposed here, not statements attributed to the providers. Sections 5.4 and 5.5 separate claims and record all three checks.

The staffing examples show why the unit and denominator need explicit attention. QS explicitly includes research staff and both degree levels in its ratio, so a university-wide figure does not identify the teaching time available in a particular undergraduate programme. Gibbs (2010, pp. 15–16) notes that staffing ratios reported at institutional level may bear little resemblance to those students experience within a department and year of study, and that, once entry characteristics are taken into account, educational gains have been found to be largely unrelated to such ratios. A more recent analysis published by the Organisation for Economic Co-operation and Development (OECD) makes the same point about rankings generally: their information on teaching relies on input measures such as staffing ratios or funding per student that often lack an evidenced relationship to quality (Van Damme, 2022, p. 25). Retaining the ratio as a resource description preserves information, while a claim about access to teachers would require programme-level allocation and contact evidence. It would be equally unwarranted to infer poor teaching from the ratio's failure to measure learning.

Doctoral awards illustrate a different boundary. Doctoral education is an educational activity, so classifying a doctoral indicator automatically as unrelated to education would be mistaken. The difficulty arises when evidence concerning doctoral provision is used to infer undergraduate effectiveness. Similarly, an institution may be highly regarded without the reputation survey constituting an assessment of students' capabilities; Gibbs (2010, p. 35) summarises evidence from the United States that reputational rankings track research funding, selectivity and expenditure, none of which predict educational gains. Van Damme (2022, p. 25) observes that rankings use academic reputation surveys to compensate for the absence of reliable teaching and learning metrics. The framework therefore distinguishes relevance to education from sufficiency for a particular educational conclusion.

Table 5 shows that domain labels do not settle the meaning of the underlying evidence. It does not establish that students draw the stronger inferences considered here or that providers encourage them; demonstrating prevalence would require a study of actual communications or user responses. Section 5.5 examines two documented communications.

### 5.3 When educational evidence does not yield a single order

Consider a second invented comparison using two educational assessments, each scored from 0 to 100 and assumed comparable across the institutions. University C scores 80 on disciplinary knowledge and 60 on applied problem-solving; University D scores 70 and 75 respectively. C leads on knowledge, while D leads on application. Neither institution exceeds the other on both measures. Separate reporting preserves the disagreement. A single order requires a choice about the relative importance of the outcomes, even though research has been removed entirely.

If both measures are treated as equally important and linearly interchangeable, their arithmetic means are 70 for C and 72.5 for D. Giving knowledge a weight of 0.70 changes the scores to 74 and 71.5. The order reverses because the valuation changes. A common numerical range does not establish that ten points on one assessment carry the same educational meaning as ten on the other. The combined scores are defensible only as illustrations under the stated assumptions; the arithmetic supplies no justification for those assumptions.

Uncertainty imposes a further limit even when the outcome is agreed. Suppose two programmes differ by two points on one comparable learning measure, with a hypothetical 95% confidence interval from  $-1$  to  $5$ . The point estimate favours one programme, but the interval includes differences in either direction. Failure to establish superiority is not proof of equivalence, which would require a justified margin and an appropriate analysis; descriptive differences, statistical precision and educational importance therefore require separate judgements.

### 5.4 Completed hypothetical claim records

The following records use invented conditions and numbers; the assumed measurement documentation is part of each scenario. In each record, every check receives one of the three statuses defined in Section 4.1: S (sufficient support), I (insufficient support: the documented evidence does not justify the wording) or U (indeterminate support: information needed to judge the wording is missing). Each status applies to the exact wording under examination.

#### Case 1: Retaining an observed learning-gain claim

Claim H1: "The mean disciplinary-knowledge score of the 80 students with matched assessments increased by eight points between entry and completion in Programme P." Stipulated evidence: a fictional assessment register records 100 entrants, 80 matched records, a mean of 52 at entry and 60 at completion for those same 80 students, and the absence of final scores for the other 20. A fictional technical note supports the interpretation of the content, score comparability across occasions, and consistent administration; the paired-score register confirms the calculation. The claim describes recorded scores, not the causal effect of attending the programme.

Construct — S: the outcome is the assessed disciplinary knowledge supported by the technical note, not overall educational quality. Scope and comparison — S: the statement identifies the matched group and period, and the within-group comparison uses a common scale and the same students without claiming anything about all 100 entrants or another university. Inference — S: the arithmetic

supports an observed mean increase, and the wording makes no causal attribution or generalisation beyond the observed group; measurement limitations still accompany the report.

Decision: retain H1 with the note that matched records cover 80 of 100 entrants. The eight-point change cannot describe the entering cohort, establish educational importance without a justified benchmark or show that Programme P caused the gain; those would be separate claims requiring separate records.

### **Case 2: Narrowing a survey claim and reassessing it**

Claim H2: "Students in Programme Q received effective feedback in their final semester." Stipulated evidence: all 100 enrolled students were invited to a final-semester questionnaire, and 60 responded, and 48 respondents agreed with the item "Feedback was easy to understand". The fictional survey file contains the wording, response options, administration date and anonymised counts, but no learning measure or non-response analysis. The item measures reported clarity, not the effect of feedback on learning.

Initial checks: construct — I, because agreement about clarity does not establish effectiveness; scope and comparison — U, because the records do not establish whether the respondents represent all students; inference — I, because the survey alone does not establish an educational effect. Decision: withhold H2 and enter the NARROW step. Proposed revision H2-R1: "In Programme Q's final-semester survey, 48 of 60 respondents agreed that feedback was easy to understand; 100 students were invited."

Reassessment of H2-R1: construct — S, because the statement reports the response actually elicited; scope and comparison — S, because the respondent group, denominator and semester are explicit and no institutional comparison is claimed; inference — S, because verified counts support the descriptive statement without attributing learning to feedback. Retain H2-R1 and preserve H2 as withheld. Non-response remains a limitation for generalisation, but does not prevent an accurate account of these responses. The revised statement preserves information about feedback while removing the unsupported claim of effectiveness.

### **Case 3: Withholding an educational comparison**

Claim H3: "University R contributes more to undergraduate learning than University S." Stipulated evidence: a fictional bibliometric record gives R a higher field-normalised citation score than S for the same publication period. Its coverage and calculation are documented, but no educational outcome, validated educational prediction, or institutional-effect evaluation is available. The citation comparison is the sole proposed support for H3.

Construct — I: the measure concerns citation influence, not undergraduate learning or institutional contribution. Scope and comparison — U: no information about undergraduate cohorts or programme conditions is available to assess their comparability; comparable publication records do not resolve that gap. Inference — I: the bibliometric difference supplies no basis for attributing a difference in learning to the institutions. Decision: withhold H3. There is no narrower educational outcome claim supported by the stipulated record. A separate statement about citation influence may be considered within the research profile; it is not a revised educational claim.

The record identifies what could reopen H3: relevant learning outcomes, defensible comparability and an evaluation capable of supporting attribution. It does not conclude that the institutions teach equally well or that either performs poorly. Unlike Case 2, no part of the original educational

information survives; a statement about research influence would be a different claim, not a narrowed one.

## 5.5 Illustrations from documented statements: a TEF rating and a ranking pillar score

The stipulated cases fix the evidence in advance; documented statements illustrate how the procedure applies when the evidence is real and the wording already in public use. Two were selected because each attaches an educational interpretation to a published rating or score. They are illustrations chosen and interpreted by the author, not a test of the framework's effectiveness. The first concerns the TEF. According to the Office for Students (2023), participating providers receive an overall rating and two aspect ratings, for student experience and student outcomes, which reflect the extent to which a provider delivers excellent experience and outcomes for its mix of undergraduate students across its courses and subjects. A Gold rating signifies that experience and outcomes are "typically outstanding", and each rating is published with the provider's TEF indicators and, where available, a student submission. Because the TEF is a regulatory rating with carefully bounded definitions, any overextension arises in how the rating is described.

### Case 4: Narrowing a documented public claim

A student-facing explainer by McCabe and Bhardwa (2023), published by Times Higher Education, glossed a Gold rating as meaning that a university had "demonstrated the highest-quality teaching standards" and was consistently outstanding. Claim H4 applies that gloss to a hypothetical provider, University U, assumed to hold a Gold overall rating: "University U has demonstrated the highest-quality teaching standards and is consistently outstanding." The evidence considered is the documented rating definition; a real application would first verify the provider's rating and could also examine its indicators.

The wording admits two readings, which the record assesses separately so that the claim is not rejected for a meaning its author may not have intended. Read categorically, H4a states that University U's provision falls in the TEF's highest category. Construct — S: Gold is the top rating, and its definition concerns student experience and outcomes in which teaching is central, although outcome measures also reflect student intake (Section 2). Scope and comparison — I: the gloss adds a claim of consistency that the Office for Students definition, which says "typically," does not confirm, and it does not specify whether consistency is meant across time, courses, or student groups. Inference — S: a panel judgement against stated criteria can support a categorical description. Read comparatively, H4b states that University U's teaching exceeds that of other providers. Construct — S, as for H4a. Scope and comparison — I: the rating is a categorical judgement shared by many providers (Office for Students, 2023) and does not order those within a category. Inference — I: "demonstrated" would then assert an established superiority that a categorical rating cannot carry. Decision: withhold H4 in both readings and enter the NARROW step, which must remove the shift from typically to consistently and any comparative implication.

Proposed revision H4-R1: "In TEF 2023, University U received a Gold overall rating, the highest of the TEF's rating categories, which the Office for Students defines as student experience and student outcomes that are typically outstanding for its mix of undergraduate students." Reassessment, conditional on verification that University U holds this rating: construct — S, because the statement reports the rating's documented object; scope and comparison — S, because the exercise, the student population and the qualifier are retained, and placing the rating in the highest category implies no ordering among Gold-rated providers; inference — S, because an expert judgement is reported as a

judgement. Retain H4-R1 once the rating is verified, reporting the two aspect ratings alongside it, because the overall rating does not show whether experience and outcomes were rated equally. H4 remains withheld.

The case illustrates two features of the procedure. Assessing the readings separately shows that the categorical reading is largely defensible and that the decisive problem is the shift from typically to consistently, which adds a claim of consistency that the definition does not confirm. The narrowed statement keeps what the rating legitimately conveys, including its place at the top of the TEF scale, while removing that addition and any comparative implication; a provider claiming superior teaching would need evidence of the kind discussed in Section 5.6. The analysis concerns the wording, not the provider or the design of the TEF.

### **Case 5: Narrowing an institutional description of a ranking pillar score**

The second documented statement concerns a global ranking directly. THE describes its Teaching pillar as the learning environment, and the pillar is built from the five indicators in Table 5: a teaching reputation survey, a student–staff ratio, two doctoral ratios and institutional income. Announcing its results in the 2026 edition, Kenyatta University (no date) reported its Teaching score and presented it as "reflecting the University's commitment to quality teaching, student engagement, and a supportive learning environment". A similar gloss appears in another announcement, in which the University of Reading (2025) attributed an improved teaching score to its commitment to educational excellence. Claim H5 restates the first gloss for an unnamed provider: "University V's THE Teaching score reflects its commitment to quality teaching, student engagement and a supportive learning environment."

Because the statement joins three attributes, the record assesses each component rather than assigning one verdict to the whole. It asks whether the score supports each claim, not whether University V holds the commitments described; the latter would require other evidence and cannot be determined from the score. Scope and comparison — S for each component: the claims, like the score, concern the institution as a whole in the 2026 edition, and none extends the score to particular programmes, student groups or other periods, or implies a comparison with other institutions. H5a, commitment to a supportive learning environment: construct — I, because the student–staff ratio and institutional income document resources rather than commitment and also reflect funding conditions; inference — I, because the composition of the score is known and it cannot distinguish commitment from circumstance. H5b, commitment to quality teaching: construct — I, because a reputation survey and resource ratios record how academics regard an institution and what resources it has, not actions directed at teaching quality; inference — I. H5c, commitment to student engagement: construct — I, because no indicator in the pillar concerns engagement; inference — I. Decision: withhold H5 as worded and enter the NARROW step.

Proposed revision H5-R1: "In the THE World University Rankings 2026, University V received a Teaching pillar score built from a teaching reputation survey, a student–staff ratio, two doctoral ratios and institutional income; its resource indicators describe part of the learning environment the university provides." Reassessment: construct — S, because the statement names what the score is built from and attributes to the resource indicators only what they document; scope and comparison — S, because it reports an institution-level score from a specified edition; inference — S, because no conclusion about intent, teaching quality or engagement is drawn. Retain H5-R1. A claim of commitment to quality teaching could be supported separately by documented policies or investment, and a claim about engagement by an engagement measure with documented coverage,

each with a record of its own. Case 5 complements Case 4: there, the problem was a scope extension within a bounded rating; here, attributes the score does not record were attached to it. In both, the procedure preserves an accurate statement, withholds only the unsupported elements and makes no judgement about the institution's teaching.

## 5.6 What survives: comparative teaching claims

Applied consistently, the checks have a consequence that the earlier sections only imply. A claim of teaching superiority, understood here as a statement that one institution produces more undergraduate learning than another, requires a common educational construct, comparable populations and measurement conditions, and a design able to separate institutional contribution from student selection. Descriptive comparisons of reported experience or final attainment require less (Table 3), but they do not support that conclusion. None of the indicators examined in Section 5.2 meets these requirements unless combined with evidence on learning. Research measures and reputation fail the construct check because they record research output and academic opinion rather than learning, and staffing ratios fail it because they describe resources. None of these indicators can separate institutional contribution from other influences, as the inference check requires. A further scope problem arises if institution-wide ratios are applied to particular programmes. The absence of any learning measure among these indicators is consistent with the OECD's observation that international rankings capture research-related and input measures but provide no insight into students' learning outcomes (Van Damme and Zahner, 2022, p. 15). Reweighting these indicators cannot remedy the deficiency, and declared weights may not even describe their effective influence (Section 5.1).

The gap does not arise because learning outcomes cannot, in principle, be compared across institutions. The OECD's Assessment of Higher Education Learning Outcomes (AHELO) feasibility study involved 248 institutions and about 23,000 students in 17 countries or economies. It concluded that such an assessment was feasible, while acknowledging substantial conceptual, methodological and practical difficulties (Van Damme, 2022, p. 30). A proposed main study nevertheless failed to gain sufficient support among member countries, amid strong concerns and opposition from leading universities and university associations; critics also argued that higher education is too diverse for common measures (Van Damme, 2022, pp. 30–31; Van Damme and Zahner, 2022, pp. 3, 15). A smaller group of countries continued through the CLA+ International initiative, which used the Collegiate Learning Assessment (CLA+) to assess critical thinking and written communication in six countries between 2016 and 2021. The editors present the results as informative rather than definitive (Van Damme and Zahner, 2022, pp. 15–16). The obstacles are therefore both technical and institutional. The technical difficulties include cultural adaptation, student motivation and the choice between absolute and value-added measurement (Van Damme, 2022, p. 31; Van Damme and Zahner, 2022, p. 16). Whatever their eventual resolution, none of the ranking indicators examined here draws on such evidence.

The framework therefore implies a division of labour rather than a prohibition. Global tables can legitimately report research performance, resources and reputation under accurate labels. The educational statements that survive are those in the middle column of Table 2 – reported experience among identified respondents, attainment on assessed tasks and observed progress on comparable measures – together with estimates of institutional contribution (Table 1) where designs permit and comparisons are confined to comparable groups. Each is bounded by what its evidence covers; broader comparisons remain possible where measures and designs support them. A global ordinal list of teaching quality would require comparable learning-outcome evidence, validated for the

intended comparison, of the kind these initiatives have begun to develop. Until such evidence is incorporated, "insufficient evidence for comparison" is the defensible educational entry, and a ranked position should not be read as a verdict on teaching.

## **6 Objections and implementation**

### **6.1 Research-informed education and institutional missions**

One objection is that separate reporting understates the intellectual unity of research and teaching. As Section 2.1 noted, the framework separates claims rather than activities; requiring educational evidence makes visible the mechanisms through which research enriches education, instead of assuming that every research output benefits every learner. A related objection concerns research-intensive institutions. Separate reporting accommodates their missions, because research performance remains visible and assessable on its own terms; the proposal does not require every university to weight teaching and research equally, only that the meaning of each assessment remains clear.

### **6.2 Educational aims and the demands of measurement**

Educational evidence can be expensive to collect and difficult to compare. Longitudinal measurement requires sustained participation, disciplinary assessment requires expertise, and contextual analysis requires appropriate data. These constraints argue for proportionate reporting and bounded comparisons rather than a universal demand for one international teaching score. A staged implementation could first clarify existing labels and disclose evidence gaps, then improve programme-level reporting from existing assessment and student-experience evidence, and finally add targeted measurement where important questions remain unanswered.

Nor should randomised experiments become a universal condition of educational accountability: descriptive and observational evidence can answer worthwhile questions, provided causal language is reserved for designs that warrant it. The choice of outcomes is itself consequential. A programme may legitimately seek professional judgement, intellectual independence or collaborative capability alongside examination performance, so assessors should explain why the selected outcomes represent its purposes and what remains outside the measures. Moderated student work, professional assessments and qualitative evidence can contribute when their methods and scope are explicit; the framework requires evidence appropriate to a claim, not the conversion of every achievement into a standardised score.

### **6.3 Responsibilities for using assessment evidence**

A provider adopting the framework should identify who approves educational claims, publish the evidence requirements before applying them and make completed records available for scrutiny. Institutions should be able to correct source data or challenge an inference, with reasons for accepting or rejecting a correction documented and a dated update issued, so that users can identify which version informed a decision.

Governing bodies and public agencies should state which decision an assessment informs, because evidence adequate for describing student experience may be inadequate for allocating funds on the basis of superior teaching; the framework supplements broader quality assurance rather than replacing it, and it does not supply a funding formula. An evidence gap should prompt a proportionate request for information rather than an automatic penalty. A pilot recording staff time, disputed classifications, correction requests and effects on institutions with limited reporting

capacity would show whether clearer reporting is feasible without disproportionate burden. Where a decision cannot wait for stronger evidence, students and agencies may act on the available evidence together with costs, access and stated priorities, provided the decision record explains the remaining uncertainty; a defensible choice under uncertainty remains distinct from a demonstrated claim of educational superiority.

The stakes are higher where rankings inform national strategy. Thailand's Reinventing University initiative, launched by the Ministry of Higher Education, Science, Research and Innovation in November 2020, placed 82 of 155 universities into strategic groups in a first round announced in November 2021: 16 in a Global and Frontier Research group expected to be globally competitive, 41 in an Area-Based and Community group oriented towards local development, and others in technology-and-innovation and specialised-professional groups, with enhancing teaching and learning quality among the activity themes (Australian Government Department of Education, Skills and Employment, no date). Research-weighted ranking positions, to which Thai universities have responded strategically (Srisawat, 2020; Section 3), say little about the educational contribution of universities whose missions are local or professional. Separate reporting with bounded educational claims would allow research ambitions to be recognised without converting them into educational verdicts, and would let teaching-focused missions be assessed on evidence suited to them.

#### **6.4 Simplicity and the use of ranking information**

A single ordinal list offers an easily recognised result, and multiple profiles impose greater demands on readers. The added demand on readers is a substantive objection to multidimensional reporting: a technically careful report can fail if its audience cannot interpret it. Evidence on presentation supports treating communication as a design problem in its own right. Luca and Smith (2011) relate the salience of rankings to application behaviour among highly ranked US institutions. Hohensinn et al. (2026) conducted an eye-tracking experiment in which students chose from university performance tables, using fixation duration as a proxy for cognitive processing. They report that rankings reduce attention to underlying information, while acknowledging that adding information may itself contribute to the redistribution of attention. Neither study tests whether participants reached accurate interpretations, and neither validates the format proposed here.

The practical implication is layered presentation: a concise public statement of each retained claim, a small number of clearly differentiated educational dimensions, and an accessible claim record for readers who need the reasoning. Whether this improves comprehension is an empirical question (Section 6.5).

#### **6.5 Conceptual adequacy and limitations**

This article assesses the framework for coherence, fit with established validity principles, explicit decision requirements, and consistency across cases. The stipulated cases show retention, narrowing and withholding under controlled conditions, and the documented cases illustrate how the procedure separates defensible elements of public statements from overextended ones. These illustrations provide a conceptual justification and an inspectable specification, not evidence that independent assessors would reach the same decisions or read the statements in the same way.

Practical performance requires empirical evaluation of agreement between independent assessors, the quality of retained and revised statements, user comprehension, and reporting burden, with attention to both unwarranted acceptance and unnecessary withholding. Implementation could also encourage emphasis on easily measured outcomes or selective reporting; disclosure of exclusions, a

record of revisions, and examination of unfavourable evidence would make such practices easier to question, but whether these safeguards work, and whether the burden falls disproportionately on less-resourced institutions, remains to be tested.

The synthesis has four further limits. It is targeted rather than systematic, so closer precedents may exist. The indicator analysis uses documented definitions rather than institutional data, and the conclusion of Section 5.6 is accordingly confined to the indicators examined; extending it to other rankings or editions requires the same analysis of their documented indicators. The documented cases illustrate wording in public use but do not estimate its prevalence. The undergraduate focus limits extension to postgraduate and doctoral education.

## 7 Conclusion

Research performance and educational claims require evidence appropriate to their different meanings. Separating their presentation helps preserve those meanings, but an educational heading does not establish learning or institutional contribution. The relevant question is what the available observations justify saying about specified students, programmes and periods.

The article has proposed a claim record that answers that question statement by statement. Each educational claim needs an identifiable construct, a population and, where applicable, a comparison, together with a justification for the inference drawn; each receives a disposition, and any narrowed version is assessed afresh and linked to the wording it replaces. Applied to documented ranking indicators and to public descriptions of a TEF rating and a ranking pillar score, the procedure preserves useful descriptions of resources, reputation, experience and attainment while withholding stronger conclusions that those descriptions do not support.

The broader implication is that keeping the meaning and evidence of each domain separately inspectable is necessary but not sufficient. Without additional evidence on student learning, the indicators examined here cannot show that one university adds more to its undergraduates' learning than another, and an education profile built on the evidence currently available is likely to be narrower in scope, more qualified and less ordinal than a league table. Users remain free to weight separately reported results as they wish; what the framework rules out is presenting the resulting decision score as demonstrated teaching effectiveness. Whether this form of reporting improves interpretation in practice is a question for empirical study; the conceptual case is that neither a label nor the numerical precision of a result should carry more educational meaning than its evidence supports.

## References

- American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 2014. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. Available at: <https://www.testingstandards.net/open-access-files.html> (Accessed: 5 October 2026)
- Australian Government Department of Education, Skills and Employment, no date. *Thailand education policy update: Reinventing University – a step towards the future of Thai higher education*. Available at: <https://www.education.gov.au/download/13189/thailand-education-policy-update-reinventing-university/25440/thailand-education-policy-update-reinventing-university/pdf> (Accessed: 5 October 2026)

- Cifuentes Gomez, G. and Santelices, M.V., 2024. 'Is it worth attending higher education? Lessons from a systematic review on institutional contribution to learning outcomes', *Cogent Education*, 11(1), 2351241. Available at: <https://doi.org/10.1080/2331186X.2024.2351241>
- De Witte, K. and Hudrlikova, L., 2013. 'What about excellence in teaching? A benevolent ranking of universities', *Scientometrics*, 96(1), pp. 337–364. Available at: <https://doi.org/10.1007/s11192-013-0971-2>
- Federkeil, G., 2022. *U-Multirank publishes new 2022 outcomes: new insights on student diversity, access and social inclusion*. CHE Centre for Higher Education, 21 June. Available at: [https://www.che.de/en/2022/u-multirank-2022\\_international/](https://www.che.de/en/2022/u-multirank-2022_international/) (Accessed: 5 October 2026)
- Gadd, E., Holmes, R. and Shearer, J., 2021. 'Developing a method for evaluating global university rankings', *Scholarly Assessment Reports*, 3(1), 2. Available at: <https://doi.org/10.29024/sar.31>
- Gibbs, G., 2010. *Dimensions of quality*. York: Higher Education Academy. Available at: <https://www.advance-he.ac.uk/knowledge-hub/dimensions-quality> (Accessed: 5 October 2026)
- Hattie, J. and Marsh, H.W., 1996. 'The relationship between research and teaching: a meta-analysis', *Review of Educational Research*, 66(4), pp. 507–542. Available at: <https://doi.org/10.3102/00346543066004507>
- Hattie, J. and Marsh, H.W., 2004. 'One journey to unravel the relationship between research and teaching', paper presented at *Research and teaching: closing the divide? An international colloquium*, Marwell Conference Centre, Winchester, 18–19 March. Available at: [https://cdn.auckland.ac.nz/assets/education/about/research/documents/relationship-between-research-and-teaching-\(2004\).pdf](https://cdn.auckland.ac.nz/assets/education/about/research/documents/relationship-between-research-and-teaching-(2004).pdf) (Accessed: 6 October 2026)
- Healey, M., 2005. 'Linking research and teaching: exploring disciplinary spaces and the role of inquiry-based learning', in Barnett, R. (ed.) *Reshaping the university: new relationships between research, scholarship and teaching*. Maidenhead: McGraw-Hill/Open University Press, pp. 67–78.
- Hohensinn, L., Willems, J., George, B. and Van de Walle, S., 2026. 'Performance rankings reduce cognitive processing of underlying performance information', *Public Management Review*, 28(2), pp. 498–520. Available at: <https://doi.org/10.1080/14719037.2025.2464761>
- Ioannidis, J.P.A., Patsopoulos, N.A., Kavvoura, F.K., Tatsioni, A., Evangelou, E., Kouri, I., Contopoulos-Ioannidis, D.G. and Liberopoulos, G., 2007. 'International ranking systems for universities and institutions: a critical appraisal', *BMC Medicine*, 5, 30. Available at: <https://doi.org/10.1186/1741-7015-5-30>
- Kane, M.T., 1990. *An argument-based approach to validation*. ACT Research Report Series 90-13. Iowa City, IA: American College Testing Program. Available at: <https://files.eric.ed.gov/fulltext/ED336428.pdf> (Accessed: 5 October 2026)
- Kane, M.T., 2013. 'Validating the interpretations and uses of test scores', *Journal of Educational Measurement*, 50(1), pp. 1–73. Available at: <https://doi.org/10.1111/jedm.12000>
- Kenya University, no date. *Kenya University shines in the Times Higher Education World University Rankings 2026*. Available at: <https://www.ku.ac.ke/times-higher-world-rankings-2026/> (Accessed: 5 October 2026)

- Luca, M. and Smith, J., 2011. *Saliency in quality disclosure: evidence from the U.S. News college rankings*. Working Paper 12-014. Boston, MA: Harvard Business School. Available at: [https://www.hbs.edu/ris/Publication%20Files/12-014\\_0e98e232-ae41-4761-974b-5b2fad7acee9.pdf](https://www.hbs.edu/ris/Publication%20Files/12-014_0e98e232-ae41-4761-974b-5b2fad7acee9.pdf) (Accessed: 5 October 2026)
- McCabe, G. and Bhardwa, S., 2023. *What is the TEF? Results of the Teaching Excellence Framework 2023*. THE Student, 18 December, updated 28 September 2026. Available at: <https://www.timeshighereducation.com/student/news/tef-2023-results> (Accessed: 5 October 2026)
- Office for Students, 2023. *TEF 2023 ratings*. Available at: <https://www.officeforstudents.org.uk/for-providers/quality-and-standards/tef-2023-ratings/> (Accessed: 5 October 2026)
- Palali, A., van Elk, R., Bolhaar, J. and Rud, I., 2017. *Are good researchers also good teachers? The relationship between research quality and teaching quality*. CPB Discussion Paper 347. The Hague: CPB Netherlands Bureau for Economic Policy Analysis. Available at: <https://www.cpb.nl/sites/default/files/omnidownload/cpb-discussion-paper-347-are-good-researchers-also-good-teachers.pdf> (Accessed: 6 October 2026)
- Palali, A., van Elk, R., Bolhaar, J. and Rud, I., 2018. 'Are good researchers also good teachers? The relationship between research quality and teaching quality', *Economics of Education Review*, 64, pp. 40–49. Available at: <https://doi.org/10.1016/j.econedurev.2018.03.011>
- Paruolo, P., Saisana, M. and Saltelli, A., 2012. 'Ratings and rankings: voodoo or science?', arXiv preprint arXiv:1104.3009v2 [stat.AP], 20 July. Available at: <https://arxiv.org/abs/1104.3009> (Accessed: 5 October 2026). Published version: *Journal of the Royal Statistical Society Series A*, 176(3), 2013, pp. 609–634. Available at: <https://doi.org/10.1111/j.1467-985X.2012.01059.x>.
- QS Quacquarelli Symonds, 2024. *Faculty-student ratio (indicator)*. Updated 28 November 2024. Available at: <https://support.qs.com/hc/en-gb/articles/360019108240> (Accessed: 5 October 2026)
- QS Quacquarelli Symonds, 2025. *Citations per faculty (indicator)*. Updated 2 June 2025. Available at: <https://support.qs.com/hc/en-gb/articles/360019107580> (Accessed: 5 October 2026)
- Rogaten, J., Rienties, B., Sharpe, R., Cross, S., Whitelock, D., Lygo-Baker, S. and Littlejohn, A., 2019. 'Reviewing affective, behavioural, and cognitive learning gains in higher education', *Assessment & Evaluation in Higher Education*, 44(3), pp. 321–337. Available at: <https://doi.org/10.1080/02602938.2018.1504277>
- Selten, F., Neylon, C., Huang, C.-K. and Groth, P., 2020. 'A longitudinal analysis of university rankings', *Quantitative Science Studies*, 1(3), pp. 1109–1135. Available at: [https://doi.org/10.1162/qss\\_a\\_00052](https://doi.org/10.1162/qss_a_00052)
- Srisawat, P., 2020. 'Global university rankings and implications for Thai universities' (in Thai), *Journal of Business, Economics and Communications*, 15(3), pp. 1–12. Available at: <https://so02.tci-thaijo.org/index.php/BECJournal/article/view/235779> (Accessed: 5 October 2026)
- Times Higher Education, 2025. *Methodology for overall and subject rankings for the Times Higher Education World University Rankings 2026*. September. Available at: [https://www.timeshighereducation.com/sites/default/files/breaking\\_news\\_files/the\\_world\\_university\\_rankings\\_2026\\_methodology.pdf](https://www.timeshighereducation.com/sites/default/files/breaking_news_files/the_world_university_rankings_2026_methodology.pdf) (Accessed: 5 October 2026)

- University of Reading, 2025. *UK top 30 for Reading: THE World University Rankings 2026*. 9 October. Available at: <https://www.reading.ac.uk/news/2025/University-News/UK-top-30-for-Reading-THE-World-University-Rankings-2026> (Accessed: 5 October 2026)
- Uttl, B., White, C.A. and Wong Gonzalez, D., 2017. 'Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related', *Studies in Educational Evaluation*, 54, pp. 22–42. Available at: <https://doi.org/10.1016/j.stueduc.2016.08.007>
- Van Damme, D., 2022. 'Do higher education students acquire the skills that matter?', in Van Damme, D. and Zahner, D. (eds.) *Does higher education teach students to think critically?* Paris: OECD Publishing, pp. 18–38. Available at: [https://www.oecd.org/en/publications/does-higher-education-teach-students-to-think-critically\\_cc9fa6aa-en/full-report/component-5.html](https://www.oecd.org/en/publications/does-higher-education-teach-students-to-think-critically_cc9fa6aa-en/full-report/component-5.html).
- Van Damme, D. and Zahner, D. (eds.), 2022. *Does higher education teach students to think critically?* Paris: OECD Publishing. Available at: <https://doi.org/10.1787/cc9fa6aa-en>
- Vermunt, J.D., Ilie, S. and Vignoles, A., 2018. 'Building the foundations for measuring learning gain in higher education: a conceptual framework and measurement instrument', *Higher Education Pedagogies*, 3(1), pp. 266–301. Available at: <https://doi.org/10.1080/23752696.2018.1484672>
- Vernon, M.M., Balas, E.A. and Momani, S., 2018. 'Are university rankings useful to improve research? A systematic review', *PLOS ONE*, 13(3), e0193762. Available at: <https://doi.org/10.1371/journal.pone.0193762>
- Waltman, L., Wouters, P. and van Eck, N.J., 2017. 'Ten principles for the responsible use of university rankings', *CWTS Blog*, 17 May. Available at: <https://www.cwts.nl/blog?article=n-r2q274> (Accessed: 5 October 2026)