NORC Research Science

FINAL REPORT
SEPTEMBER 2025

International College Ranking Systems: A Methodological Review

Presented by:
NORC at the University of Chicago
Soubhik Barari, Ph.D.
Ji Eun Park, Ph.D.
Susan Paddock, Ph.D.

Authors also thank colleagues from NORC, Norman Bradburn and Debra Stewart who provided insight and expertise for this report.

Presented to:
Vanderbilt University
Darren Reisberg, J.D.
Olivia Kew-Fickus, M.B.A.

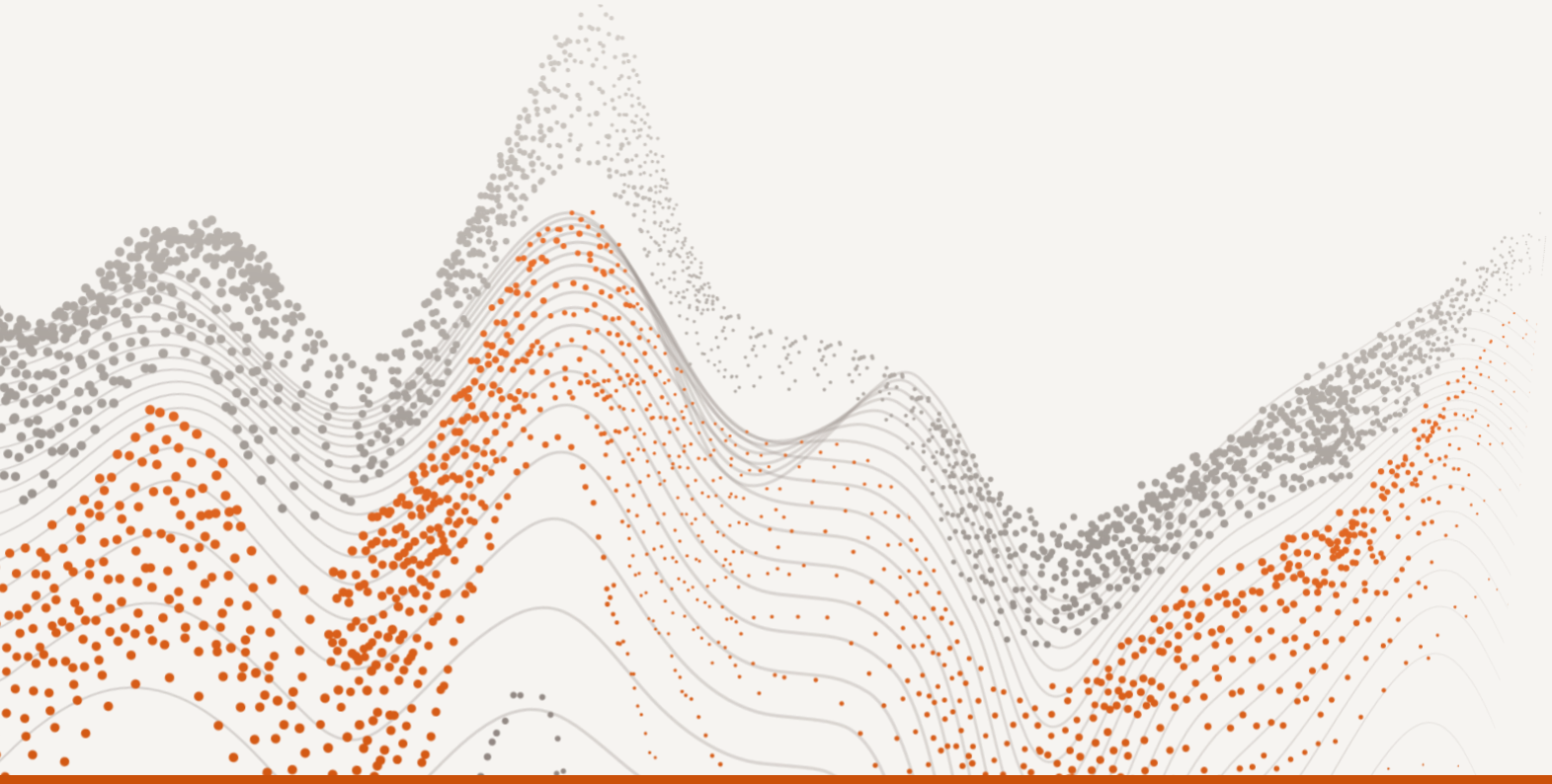

Table of Contents

About NORC

NORC at the University of Chicago is an independent 501c3 non-profit organization that conducts objective, nonpartisan research and delivers insights that decision-makers can trust. While NORC is an affiliate of the University of Chicago (UChicago) it maintains autonomy in its research activities. The University of Chicago did not provide any input on this report, and the findings and conclusions presented are solely those of NORC.

Executive Summary

Rankings of international higher education institutions have become influential, newsworthy metrics worldwide. This is unsurprising given the rising prevalence of cross-national college attendance. This report evaluates the methodologies of three major international ranking systems: the Academic Ranking of World Universities (ARWU, or "Shanghai Rankings"), the Times Higher Education World University Rankings (THE), and the Quacquarelli Symonds World University Rankings (QS).

This report serves as an addendum to our prior report (Barari et al., 2024), for which we employed a construct validity framework in reviewing the methodology employed primarily in U.S. college rankings (using *U.S. News & World Report* (USNWR), The *Wall Street Journal* (WSJ) and *Forbes* national rankings as specific examples), but also included relevant assessments of THE and QS. Construct validity is a crucial property in the development of numerical scales, such as rankings, that aim to quantitatively describe otherwise abstract qualities of social entities such as 'market value' or 'educational quality.' Our assessment included a review of conceptual, data, and methodological factors within this framework that are central to developing a methodologically sound ranking of higher educational institutions. We found four key issues spanning the conceptual, data, and methodological aspects of construct validity in that review – namely:

- There is not one clear or stable set of concepts underlying the college ranking systems
- Data quality issues limit the ability to capture even well-defined concepts
- Many methodological procedures involve subjective decisions
- There is insufficient characterization of uncertainty in numeric rankings.

We find that issues we identified with domestic rankings in the U.S. also apply to international systems, with some additional challenges particular to the international case. To summarize:

1. ***There are inherent challenges to cross-national standardization and comparability.*** Standardizing and comparing institutions across diverse cultural and educational contexts are inherently difficult tasks. Perceptions of quality can differ across cultural contexts, while data quality and reporting standards vary significantly between countries. These disparities exacerbate issues such as scale reliability and measurement invariance, often requiring arbitrary thresholds for inclusion, which can lead to selection bias. Not only does this affect the construction of the rankings, but it makes comprehensive psychometric validations of the final measures potentially impossible, posing a barrier to construct validity.

2. ***There are many layers of subjectivity in international rankings.*** In particular, the heavy emphasis on research outputs and the use of arbitrary thresholds in their methodologies involve value judgments without consultation of external guiding criteria or clear, transparent documentation to justify the decisions made. Subjectivity is particularly an issue in reputation surveys, which—while useful in capturing experts' perceptions of institutional quality—are vulnerable to cultural biases and incentive-based distortions. Such subjective evaluations are self-reinforcing or subject to anchoring effects: respondents' perceptions of institutional reputation may be heavily or entirely influenced by prior subjective measures of institutional

reputation. Consequently, rankings both measure and influence reputation. In Section 5, we propose several ways to reduce these biases in reputation surveys.

3. ***The intended purpose of each ranking is not clear.*** Our evaluation ultimately rests on the intended purpose of each ranking along these lines – as we show in the Introduction, this is not entirely clear from the stated goals of the providers. If international rankings intend to comprehensively capture student preferences, the current emphasis on prestige and research may be crowding out other necessary dimensions of interest such as teaching quality**.** For ranking consumers, research output may only serve as a proxy for more important factors that directly impact students, as evidenced by surveys of student priorities (McMurtrie, 2023). Further, research output measures based on bibliometric data are susceptible to statistical biases and require normalization to obtain comparability across universities, academic disciplines, and individual researchers. To fully capture the complexity of research productivity, multiple indices and dimensions of research performance must be employed, as we describe further in Section 5. If, however, the purpose of international rankings is to monitor institutions or inform students on a narrower set of criteria, this expansion may be less of a concern.

Issues notwithstanding, we note efforts towards robust measurement and transparency, particularly in the methodologies of the QS World University Rankings and the Times Higher Education rankings. Both systems provide extensive documentation explaining their ranking procedures and QS's methodological reports even identify and discuss many of the challenges we address in this report. Such practices should be consistently applied across ranking areas and replicated across the three systems. ARWU draws mostly from publicly available data sources and provides publicly available methodology reports, although it has been criticized for omission of methodological details, resulting in irreproducibility.

We offer several key recommendations to address these challenges which include: ***clarifying the purpose of international rankings***, which may or may not include representing the multi-faceted preferences of students beyond reputation and research quality; ***personalizing the rankings to account for differing preferences***, rather than assuming a fixed set of weights for measured categories (e.g., research outputs, reputation) based on the value judgments of experts; ***incorporating high quality data and methodology from exemplary domestic reporting systems*****,** drawing on practices at institutions or countries where data quality and reporting standards are strong; ***building more robust measures***, reducing biases from cross-cultural comparisons, citations data, and subjective reputational assessments by considering best practices in survey methodology and bibliometrics; and ***communicating when differences between institutions are meaningful***, through the usage of uncertainty statements such as confidence intervals and/or selecting the type of overall measure that accounts for this (e.g., tiers rather than ranks).

1. Introduction

University rankings have become influential, newsworthy metrics around the world. Their popularity is perhaps no surprise, given the increasing prevalence of cross-national college attendance: the rate of U.S. student enrollment in university programs[1] abroad, for instance, has more than tripled in the last thirty years (Institute of International Education, 2023).

The role of college ranking systems extends beyond media attention and student interest into policy formulation and resource allocation. At the international level, rankings influence policies affecting student mobility and the global standing of higher education institutions. Countries and immigration authorities often rely on these rankings to determine visa policies, recognizing ranked institutions as legitimate destinations for international students (Marginson, 2014). International rankings also shape rules for global competition, motivating countries to advance their higher education systems' global standings. National policies to increase research funding, foster international collaborations, or improve academic quality often aim to boost institutions' positions in global rankings, thereby enhancing the country's reputation as a leader in higher education (Altbach, 2012; 2015).

Like their American counterparts, international college rankings face a variety of methodological challenges. This report assesses the methodologies of three prominent international ranking systems: the Academic Ranking of World Universities (ARWU, also known as the "Shanghai Rankings"), the Times Higher Education World University Rankings (THE), and the Quacquarelli Symonds World University Rankings (QS). To date, these rankings are the longest running among global rankings of higher education institutions and are widely used by institutions, students, and policymakers worldwide (Marginson, 2014; Bekhradnia, 2016). Though the systems have their distinct methodologies, they share many features in common: ARWU focuses primarily on bibliometric data to measure institutional performance, THE combines bibliometric data with global reputation surveys, while QS balances different 'lenses' of higher-educational institutions which include academic reputation and employer perceptions alongside metrics like faculty-to-student ratios and research citations.

In their own words, the goals of the ranking systems (in their most recent iteration at the time of writing[2]) are as follows:

- ARWU: "Presenting the world's top universities annually based on transparent methodology and objective third-party data" … "recognized as the precursor of global university rankings and the most trustworthy one." (ShanghaiRanking, 2024)

- THE: "Judge research-intensive universities across all their core missions: teaching, research, knowledge transfer and international outlook" with significant methodological updates in 2024

[1] In this report, we use "college" interchangeably with "undergraduate institution," reflecting its usage in the U.S. context. We note that in many countries, the term "college" carries a different meaning (e.g. shorter, career-oriented programs or secondary education more broadly).

[2] In this report, we evaluated the most current iteration of the three ranking systems as of February 2025. Since we began writing, QS and ARWU published its updated rankings (for the 2026 ranking cycle) in August 2025. In our review, this includes changes to the exact indicators included in each category (e.g., QS now includes international student diversity in its measure of global engagement), methodological decisions such as inclusion criteria and weights (e.g., cut-offs), and the processing of measures (e.g., normalization). In many cases, these updates improve on elements discussed in this report, we do not believe they represent a major departure from the previous cycle's methodology.

"so that it continues to reflect the outputs of the diverse range of research-intensive universities across the world." (Times Higher Education, 2024)

- QS: "Shines a light on the best institutions from across the world, supporting our mission of enabling motivated people anywhere in the world to fulfill their potential through educational achievement, international mobility, and career development." (Quacquarelli Symonds, 2024)

Using these and other publicly available statements about each methodology (see Appendix Table A0), we evaluate the three systems on the basis of their construct validity following from our prior report (Barari et al., 2024). Construct validity is a crucial property in the development of numerical scales, such as rankings, that aim to quantitatively describe otherwise abstract qualities of social entities such as 'market value' or 'educational quality.' Barari et al. (2024) included a review of conceptual, data, and methodological factors within this framework that are central to developing a methodologically sound ranking of higher educational institutions. We found four key issues spanning the conceptual, data, and methodological aspects of construct validity in that review – namely:

- There is not one clear or stable set of concepts underlying the college ranking systems
- Data quality issues limit the ability to capture even well-defined concepts
- Many methodological procedures involve subjective decisions
- There is insufficient characterization of uncertainty in numeric rankings.

We find that the same construct validity issues affecting domestic rankings are present in international ones, though the severity varies across ranking systems and additional issues are posed by the international scope. As others have noted (Bekhradnia, 2016), the overwhelming focus on research output – and the relative lack of emphasis on teaching – creates the effect of *unidimensionality*, or the impression that only a single dimension matters when measuring the quality of the institution (though, we also note that it is nowhere stated in any system what is the particular construct of interest, 'quality' or otherwise). While the ranking providers themselves sometimes acknowledge that research output serves as a proxy for broader institutional quality, this focus likely fails to fully capture the dimensions students demonstrably care about most, such as teaching quality (Bok, 1986; Siow, 1997; McMurtrie, 2023). Moreover, the heavy reliance on reputation surveys and bibliometric data introduces biases unique to international rankings. Although some systems attempt to correct these biases, the efforts may be insufficient to address the complexities of cross-national and cross-institutional comparisons.

In short, international ranking providers like ARWU, THE, and QS must decide exactly what it is they are trying to measure, which is not entirely clear from their stated goals above. Do they intend to narrowly capture research quality and reputation, with the intention of monitoring and improving both? Or do they aim to capture how well universities deliver on the preferences of students, which may extend to criteria beyond reputation and research quality? As we reiterate in the Conclusion, our ultimate evaluation of international college rankings requires a better understanding of their intended purpose.

2. Construct Validity

We begin with a broad assessment of construct validity in the most recent (at the time of writing) QS, ARWU, and THE rankings, expanding on how the elements discussed in Barari et al. (2024) apply to international college rankings in particular. As we have established in prior assessment, achieving construct validity in any social scientific measurement requires addressing three fundamental components (Bradburn, Cartwright, and Fuller, 2017):

- **Characterization:** Define the construct, including defining the conceptual dimensions that together compose the construct, if it is a multi-dimensional construct, and identifying the boundaries of these concepts and fixing which features belong to it and which do not.

- **Representation:** Define the measurement, or the process of assigning a number to each instantiation of the construct (e.g., value) based on observable real-world features (e.g., data), that appropriately and fully describes the construct.

- **Operationalization:** Define the precise methodologies to transform the relevant real-world observations into numbers (e.g., intermediate quantities such as graduation rates) to precisely and accurately produce the final measure.

Establishing construct validity necessitates explaining how these requirements are met and harmonized. To demonstrate that characterization, representation, and operationalization of the construct align, certain key properties must be satisfied including content validity, criterion validity, and scale reliability.

Efforts toward robust measurement and transparency are evident, particularly in the most recent methodologies of QS and THE. Both THE and QS, for instance, provide extensive documentation detailing how they construct their rankings. QS is especially commendable for its transparency, offering thorough explanations of its methodology, for example describing its multi-round review process for methodological changes involving a global board of 40 advisors. As noted in our prior report, methodological changes confound the yearly college rankings: it is difficult to know whether an institution's rank changes because the *institution* changed or the *methodology* changed. To this end, QS states that experts assess and evaluate the effects of any methodological changes on the rankings.

Additionally, when individual measures deviate from regional averages or show unexpected year-to-year shifts, QS seeks supplementary information from institutions and cross-references data with national databases like the Higher Education Statistics Agency (HESA) in the United Kingdom or the Integrated Postsecondary Education Data System (IPEDS) in the United States. THE similarly describes, in detail, the transformations and normalizations applied at each stage in its methodology reports, many of which (as we later describe) are to improve comparability. These reporting efforts by the two systems demonstrate a commitment to clarity and accuracy.

However, in our assessment, many of the significant issues faced by domestic college rankings on these fronts still affect international ranking systems, and there is a need for further documentation in order to assess construct validity. We expand on each of these issues below.

2.1. Ambiguous Constructs

The constructs underpinning ARWU, THE, and QS rankings, respectively, are ambiguously defined, which undermines construct validity. Though the ranking providers state their goals (restated in our introduction), they do not fully answer the question: *what is this ranking exactly measuring*?

For example, although ARWU aims to highlight top universities based on "transparent methodology and objective third-party data," this goal does not clearly define which aspects of 'quality' are being measured, nor does it specify the boundaries of what constitutes "top" status beyond research output. THE, meanwhile, evaluates "research-intensive universities" across teaching, research, knowledge transfer, and international outlook but provides no clear conceptualization of these dimensions or how they relate to educational quality holistically. QS, similarly, promotes universities that support "educational achievement, international mobility, and career development," yet does not articulate the specific aspects of institutional quality that reflect these objectives or how they are differentiated from other constructs of university quality. This lack of precise construct definition (characterization) limits each system's ability to ensure that representation and operationalization decisions accurately capture the intended concepts.

2.2. Inconsistent Definitions of Undergraduate Institutions

First, there continue to be differing definitions of what constitutes a 'college' or 'university' across ranking systems, threatening the characterization of the ranking construct. These discrepancies arise from varying inclusion and exclusion criteria – enumerated in Table A1 – echoing issues identified in domestic rankings (Barari et al., 2024).

ARWU, for example, initially selects universities for ranking on the basis of prestigious awards and research output, the very measures used in the ranking itself. THE and QS define their respective universes as the union of a set of several criteria which suffer from the problem of *arbitrary cutoffs*. For instance, QS eligibility is contingent on the provision of full degree programs in at least 2 out of 5 broad faculty areas. Times Higher Ed's research output criterion is perhaps the most specific seen in our three systems: faculty in an eligible university "are required to publish more than 1,000 relevant publications over the previous 5 years, and more than 150 relevant publications in any single year." Circular criteria and arbitrary thresholds in the ranking systems may make the results highly sensitive to the inclusion or exclusion of specific universities. In some cases, the exclusion or inclusion of a small number of universities may significantly affect the positioning of many institutions, where previously lower-ranked universities surpass higher-ranked ones, or vice versa – one of many possible instances of *rank reversal* (Saaty, 1987). For example, for QS, Asian, Latin American, Middle Eastern, and African universities comprised 37% of this ranking system in 2018, which increased to 46% in 2024. The average ranking position of institutions from these regions also increased by 48 places among the world's top 500 between 2018 and 2024.

Notably, in some cases and with regard to some factors, care has been taken. For example, QS effectively distinguishes between institutions that choose to *opt out* of ranking participation and those that are *excluded* based on specific eligibility criteria. Institutions may be absent from a ranking due to ineligibility, often stemming from factors such as limited subject offerings, insufficient size, or failure to meet established performance metrics. Conversely, 'opting out' applies to institutions that are eligible

but do not cooperate with institutional disclosure requests for specific reasons, such as undergoing a merger with another institution or out of performance concerns. In its 2024 methodology documentation, QS clearly notes that the ranking process proceeds with or without the cooperation of individual institutions.

Still, there appears to be an element of subjective judgment in determining which institutions are included in the rankings. ARWU, for example, only includes universities with a "significant" number of papers based on bibliometric measures. To take another example, as highlighted in Table A1, QS mentions in separate documentation that new entrants to the ranking system – those previously deemed ineligible or excluded – must provide explicit justification for their inclusion. This justification is then reviewed and subject to consideration. While this additional layer of assessment can be useful for vetting fraudulent universities or institutions, it may also introduce the possibility of bias. Altogether, non-transparent and shifting decision rules in the construction of the ranking universe make it impossible to replicate the list of eligible institutions for each ranking system.

2.3. Issues with Cross-National Comparability

One reason for such particular definitions of 'college' in international rankings is to ensure greater comparability of institutions *within* a ranking list. In construct validity terms, the underlying construct being measured might not be applicable to all institutions if definitions are overly broad. For example, if the construct is defined as the "breadth of knowledge attained by students," this may not apply to trade schools or other specialized programs, such as those in Germany or Switzerland, where the primary aim is not to impart broad knowledge across subjects but to develop specialized skills in a particular area of study.

Even with standardized, narrower definitions, international rankings still require capturing the same measures for institutions that operate under vastly different educational structures, admission standards, and program lengths. These variations can complicate evaluations and lead to potentially misleading conclusions about educational quality when compared internationally.

Admission criteria, for instance, vary widely across countries, influencing the academic preparation that students bring into undergraduate programs. In many European countries, secondary education often includes specialized tracks or preparatory programs that equip students for immediate entry into focused undergraduate degrees. In contrast, U.S. students generally enter college with a broader, less specialized high school education that does not directly prepare them for a single discipline or profession.

Program length is another divergence, with U.S. undergraduate programs generally lasting four years, while those in the U.K. span only three years. Consequently, a three-year British degree may offer a comparable depth of specialized knowledge to that of a four-year U.S. degree in the same field, despite the difference in program duration.

To our knowledge, few if any measures in the ranking systems account for these fundamental differences in the structure of undergraduate education across countries. It is likely for this reason that research productivity – for which measures needed for fair comparisons are readily available (e.g., citations, institutional income, patents, research staff, field of study) – are a primary focus of international rankings rather than teaching quality. This would not be problematic should the ultimate

goal of international rankings be to assess research quality (and research quality alone) and this goal be widely understood among all stakeholders. At present, however, we cannot conclude that this is true.

2.4. Issues with Content Validity

Content validity requires that the chosen representation *fully* covers the construct, with no significant dimension of the construct omitted. In the context of college rankings, content validity means that the rankings comprehensively assess all relevant aspects of the purported construct. For example, if the construct underlying a particular ranking is 'quality,' this would require that the measurement include factors such as teaching effectiveness, research output, student satisfaction, campus facilities, and post-graduation outcomes.

The evaluation of only a few dimensions, such as research output and reputation, which do not fully capture the multifaceted nature of educational quality (or related constructs such as 'value') limits the rankings' ability to comprehensively represent meaningful constructs. This has been one of the main criticisms against the ARWU, which measures academic performance only based on research excellence (Dehon et al, 2010) as measured by research output in high impact journals and quality of faculty and education by winning of Nobel Prizes and Field Medals. The absence of key dimensions of 'quality' (according to key stakeholders such as students, as we explain further in Section 4), such as teaching quality or student satisfaction, prevent the rankings from fully satisfying content validity.

ARWU has also faced criticism for methodological bias toward science and technology disciplines, while underrepresenting fields such as the arts and humanities. Additionally, it has been criticized for favoring English-speaking universities, as English dominates academic publications (Van Raan, 2005). This bias stems from limited coverage of factors that could better assess the quality of these universities, such as incorporating non-English language data sources or using evaluation measures like literary awards or other outputs centered on the humanities.

2.5. Issues with Criterion Validity

Criterion validity (or predictive validity) requires that the final output measures correlate with or are predictive of substantively related external criterion or outcomes. For example, criterion validity may involve demonstrating that a ranking meaningfully predicts or is associated with important outcomes, such as graduate employment rates, alumni earnings, or further educational attainment.

Establishing criterion validity for international rankings is challenging due to difficulties in collecting standardized data on student outcomes or experiences across different countries and educational systems. Different countries may use varied grading systems, academic calendars, or even definitions of key outcomes like employment or graduate success, complicating cross-national comparisons. Furthermore, cultural differences in student experiences and institutional priorities add another layer of complexity, making it difficult to ensure that rankings are measuring the same attributes across diverse educational contexts. Without a common benchmark, it is difficult to assess whether the rankings accurately reflect the construct they intend to measure.

2.6. Issues with Measurement Invariance

Measurement invariance requires that the measurement operates the same way across different groups and contexts. In other words, the validity of the construct is similar across different groups. To use an example from the US domestic rankings, measurement invariance would mean that there should be similar levels of scale reliability or predictive validity for institutions regardless of size, funding, region, or academic selectivity.

In international rankings, differences in regional reporting standards and educational practices across the many regional contexts covered by international rankings may severely affect measurement invariance. For example, universities in the UK may have comparatively robust reporting mechanisms due to enforcement from the Higher Education Statistics Agency (HESA), which is mandated to collect detailed and standardized data annually from all higher education institutions. In effect, the higher level of data quality and availability may lead to UK universities being overrepresented in rankings – not necessarily because they outperform global peers, but because their data are more accessible and reliable. In some cases, the effect may also work in the reverse direction, where an abundance of high-quality data on detailed and standardized metrics determined by the national reporting requirements could lead these universities to be penalized in those dimensions.

There is a trade-off between measurement invariance and broad representation in the rankings. Measurement invariance may be improved by refining the selection criteria used to form the rankings universe in the first place to ensure a baseline degree of data quality and availability for measures used to construct ranks, but refined selection criteria might also introduce *selection bias,* leaving institutions from regions with less stringent data collection processes or reporting standards underrepresented.

2.7. Issues with Scale Reliability

Scale reliability requires demonstrating that performing the measurement process (i.e., operationalization) yields the same measures under consistent conditions. Methods such as internal consistency (via Cronbach's alpha), test-retest reliability, and inter-rater reliability can be used to assess this (Nunnally and Bernstein, 1994). Although not identical to scale reliability, the *reproducibility* (or *replicability*) of a scale under a possibly different set of external conditions at a future point in time can furnish evidence in favor of scale reliability.

Some international ranking systems have been found to lack reproducibility, though this claim has been the subject of contention. For example, despite the general accessibility of the primary data and its methodologies being publicly outlined, Florian (2007) found that ARWU's rankings could not be reproduced. Specifically, the author found that the size indicators (number of full-time equivalent academic staff) used to normalize the scores of five indicators were sometimes difficult to obtain or were inconsistent. More importantly, even when using an objective measure such as the Science Citation Index (SCI), which measures "the total number of articles indexed in Science Citation Index-Expanded, Social Science Citation Index, and Arts & Humanities Citation Index," the ranking was not reproducible with the raw data because the ranking did not follow the methodology stated in their official methodology report. Years later, Docampo (2013) argued that ARWU's results were, in fact, reproducible but the details of the publicly available methodology were ambiguous, making it difficult to replicate

exact methodological decisions – including cleaning and manipulation of the dataset – to arrive at the published ranking (Docampo 2013, Docampo et al., 2022).[3]

Furthermore, frequent changes in methodology, as seen with the Times Higher Education rankings, can lead to volatility and undermine reliability (Fidler and Parsons, 2008). Similarly, evidence suggests that ARWU's methodological changes in identifying Highly Cited Researchers (HCR) and mappings used for weighting the indicator, as well as undocumented rules, rather than intrinsic changes in the performance, have driven changes in the ARWU over the years (2004-2016) (Docampo et al., 2022). Notably, these reliability issues have been observed in other higher ed rankings such as the USNWR rankings and Financial Times' rankings of MBA programs (Iacobucci, 2013), and at the time of writing we could find no such evaluations affirming the scale reliability of the QS or THE rankings.

2.8. Problematic Proxies

Ranking systems, domestic or international, frequently must rely on *proxy variables*, or measures that serve as indirect indicators of a concept that is otherwise difficult to measure directly or unavailable. Such variables include peer reputation as measures of institutional prestige (see Section 3), bibliometric scores as measures of research quality (see Section 4), and perhaps most problematically, institutional income.

Income – most prominently featured in the THE ranking – is a potentially problematic proxy for measuring institutional quality, intended to "indicate an institution's general status" and give a "broad sense of the infrastructure and facilities available to students and staff." By their own admission, income serves only to suggest, rather than confirm the availability of such resources, which themselves do not capture the quality of education or the academic experience. Research funding, in particular, often depends on external factors unrelated to the institution's intrinsic quality, such as access to government grants or private donations which may disproportionately benefit institutions in wealthier countries or those with stronger political or corporate ties. Indeed, Times Higher Ed concedes that research income is "a somewhat controversial indicator because it can be influenced by national policy and economic circumstances." It is for these reasons that nearly all other rankings avoid the usage of income in their ranking measures.

One rationale for the use of income measures is the availability of financial data and the ease of converting income into comparable figures across regions using currency conversion rates. This approach allows for cross-regional comparisons. However, using aggregate financial measures may not be a useful proxy for constructs that actually matter for higher-ed stakeholders. In many cases, increases in financial resources or spending patterns do not even track with investments in areas that are most important to students. Recent analyses from the *Chronicle of Higher Education* show, for instance, that among domestic institutions' rising expenditures from 2005 to 2021, spending on instruction – a highly valued aspect of the collegiate experience among students – lagged compared to other categories such as student support (McMurtrie, 2023).

As we note, the issue of proxy measurement applies not only to income but to bibliometric and reputational measures as well. Additionally, measures that 'stand in' for teaching quality such as faculty/student ratio used in the QS rankings have been shown to only weakly correlate with learning

[3] See also Docampo (2011) and Docampo (2014) for further analysis of the structure and reproducibility of the ARWU rankings.

outcomes and student satisfaction in a variety of contexts compared to, for instance, teaching culture, course structure, and faculty autonomy (Light, 2001; Kuh et al., 2011; Bowen and Tobin, 2015).

2.9. The Equal Intervals Problem

Inherent issues with rank-based measures persist with international rankings, for instance, what we dubbed in earlier work (Barari et al., 2024) the 'equal intervals problem:' Ordinal rankings are easy to misinterpret. For instance, the gap between colleges ranked #2 and #3 might not be equivalent to the gap between colleges ranked #10 and #11. However, ordinal scales only indicate the order of items, not the magnitude of difference between them.

This misconception presents a fundamental issue of construct validity in college rankings. By not conveying the actual differences in the measure of the underlying construct (before it is ranked) rankings can mislead consumers. They may overestimate or underestimate how differentiated colleges truly are, potentially influencing important decisions based on an inaccurate perception of quality disparities. As we have shown in our previous report, this problem affects international rankings and domestic U.S. rankings alike.

2.10. Subjective Weights

In all cases, the scores used to construct rankings are weighted combinations of different component measures. In the ARWU and THE rankings, these weights are assigned based on judgments of internal experts, rather than on empirical or other plausibly objective criteria. In the case of the QS World University Rankings, weights are documented to be driven by surveyed student priorities; however, we could find little supporting detail about the specific methodology. In particular, it is unclear how these student priorities were aggregated (e.g., average, median, majority vote) and translated into specific weights. Elsewhere, the 2024 THE rankings used "pre-weighted" individual indicators, suggesting the removal of researcher discretion in combining lower-level subcategory measures to form the category measures, though no further detail is provided on how or why this is done.

A further issue to consider with international rankings, as compared to the domestic college rankings, is the implications of heavily weighting one component. For instance, QS allocates 45% of its weight to reputation-related category measures (Appendix Table A2). This means that one dimension of the construct is disproportionately represented in the final measure, leading to a possibly incoherent interpretation of the construct if other important dimensions are omitted or not weighted appropriately to reflect user priorities. Moreover, this provides a clear incentive for universities to optimize (alternatively, 'game' or report statistics that appear to optimize) on one particular dimension of evaluation. One such method may be the strategic allocation of resources: for instance, it is reported that Malaysian institutions actively seek and incentivize engagement on academic reputation surveys (Calderon, 2020). While 'over-engagement' itself may not necessarily corrupt the resulting measures, if institutions selectively invest in the collection (and, thus, statistical significance) of data for measures that are either heavily weighted or are expected to be 'rank-enhancing', ranks are likely to be biased.

2.11. Categorization and Transformation Choices

Beyond the use of subjectively determined weights, college ranking providers make other subjective decisions in transforming raw data into operational metrics. While subjectivity does not inherently compromise the integrity of the measurement, it introduces an additional layer of discretion, allowing providers to make choices that may favor certain preferred outcomes.

In some cases, these decisions, even if based on subjective judgments, result in improved measurement. For example, as we elaborate in Section 5, normalizing citations by subject area helps ensure fair comparisons across different fields (Cronin and Sugimoto, 2014). However, it is crucial to account for the date a paper was published, otherwise measures may bias in favor of older and larger universities (Cronin and Sugimoto, 2014). This bias would occur because older universities and larger institutions have had more time and resources to produce a greater number of publications and accumulate citations. Fortunately, we find that the various rankings normalize to fixed (though not always uniform) citation windows; ARWU, for instance, uses a citation window of 10 years for certain measures. Beyond research, the cross-national divergences in undergraduate program structures, discussed earlier, would benefit from the statistical normalization of relevant institutional measures for program length, admission standards, and curriculum structure – though this would likely necessitate substantially greater data collection efforts.

Normalization is a ubiquitous feature in ranking systems. For example, each year, ARWU assigns the top university a fixed score of 100, which serves as the benchmark against which all other universities are measured. Their scores are calculated as a proportion of this top score. This method can create the misleading impression that all universities' scores are based on some fixed and stable rubric, where 100 represents an absolute ceiling for academic performance. However, in reality, the score of 100 is relative to the highest performer for that year, and this top score itself can vary over time or even within a single year due to uncertainty in the data or issues related to scale reliability. As a result, the 100-point benchmark does not represent a consistent standard year over year, but rather a moving target that depends on the performance of the leading university at any given time. This variability introduces instability in the ranking system and further complicates interpretations of differences between institutions. Consequently, universities' scores, and their relative positions in the rankings, may reflect not only their performance but also fluctuations in the top university's score, making it difficult to assess true long-term trends or differences. This approach can distort perceptions of the quality gap between universities, amplifying the "equal intervals problem" by suggesting that differences between ranked universities are more stable and consistent than they truly are.

2.12. Communicating Uncertainty

There is inherent uncertainty about the exact position of each college on the ranking scale, similar to the case for any statistical measure. Furthermore, there is uncertainty associated with nearly every measure used in constructing each rank, such as uncertainties arising from drawing a sample of university administrators, faculty, or students for a survey. However, none of the major ranking providers communicate this uncertainty through the usage of statistical confidence intervals for all rank positions or otherwise.

It is noteworthy that QS chooses to truncate or group universities below a certain rank into tiers, which is commendable as it may be intended to convey this uncertainty. Nevertheless, this tiering approach may also be necessary for the top-ranked universities, as slight differences in their scores may not be statistically significant such that presenting them with precise rankings could be misleading.

While the aforementioned concerns afflict the rankings of both U.S. and international universities, two key features (and their associated issues) distinguish international rankings in particular: cross-national reputation surveys and bibliometric measures of faculty productivity. We now turn to a closer examination of these two components.

3. Usage of Cross-National Reputation Surveys

Compared to rankings of U.S. colleges, international ranking systems heavily emphasize *reputation*, or the perception of institutional quality from various stakeholders, including academics, employers, and students (see Tables A2-A4 for definitions and justifications). For many stakeholders, including students (Bok, 1986; Siow, 1997), reputation serves as a proxy for long-term performance, prestige, and societal impact, and it can reflect dimensions of an institution that are not easily quantified through objective data alone (Hazelkorn, 2015). For instance, academic reputation often reflects peer assessments of research quality, teaching effectiveness, and contributions to the broader academic community (Marginson, 2007).

Reputation surveys are a useful method for measuring this intangible construct because they require gathering expert opinions from individuals who have knowledge of the institutions being evaluated. Through the administration of a fixed instrument with consistent questions to the same universities, surveys allow for standardized data collection across a broad range of institutions (more than 7,000 in the case of QS). This approach allows for the aggregation of (potentially) informed judgments across a broad spectrum of respondents. Moreover, surveys can be constructed to include a range of attributes, enabling respondents to evaluate specific areas such as research output, teaching, and community engagement (Fombrun, 1996). Unlike other available sources of observational data such as enrollment or full-time faculty counts from third parties or universities themselves (each with their own incentives for non-disclosure or misreporting), surveys offer researchers the opportunity to proactively *design* data collection instruments with psychometric validity for important survey-based measures from the start. The field of survey methodology offers many tools for targeting and minimizing different types of error involving, but not just limited to, construct validity (e.g., see Groves et al. (2009) for an overview of the *Total Survey Error* framework).

In sum, reputation surveys can be flexibly designed to complement other ‘objective’ or performance-based data, offering a more holistic understanding of an institution’s global or regional standing. Moreover, reputation may in and of itself be sought after by some students – learning outcomes may be secondary to going home with a prestigious degree from a reputable university. In a 2022 survey of 420 international students enrolled across 24 U.S. colleges, a majority cited reputation as a reason for selecting the U.S. as a destination for their undergraduate studies (Obst and Forster, 2022). Nevertheless,

reputation surveys are prone to various biases. Here we discuss issues with the theoretical concept and measurement of reputation.

3.1. Cultural Biases

Reputation is often shaped by cultural norms and values, which can introduce biases in how institutions are perceived. For example, social psychology research has demonstrated that *individualistic cultures*, such as those prevalent in the United States and Western Europe, tend to emphasize personal achievement, autonomy, and self-promotion. In these cultures, reputation is often associated with individual accomplishments and institutional prestige built upon measurable successes like research output and global rankings (Markus and Kitayama, 1991; Triandis, 1995). Respondents from these backgrounds may rate universities based on well-publicized achievements, international visibility, and competitive standings.

In contrast, *collectivist cultures*, common in many Asian countries, prioritize group harmony, social relationships, and community. In these contexts, reputation may be closely tied to an institution's contributions to societal well-being, adherence to cultural norms, and the promotion of collective values (Hofstede, 2001; Heine, 2001). The tendency toward modesty and humility in some Asian cultures can result in underreporting of an institution's strengths, a phenomenon known as the *modesty bias* (Chen, Lee, and Stevenson, 1995). This bias contrasts with the *self-enhancement bias* often observed in Western cultures, where individuals may overstate their achievements (Heine and Hamamura, 2007). As we elaborate further in Section 3, such differences can skew reputation survey results, as institutions from cultures with modesty biases may receive lower reputation scores not due to lower quality but because of cultural norms influencing self-reporting. In the QS ranking methodology, this is partially ameliorated by the fact that international votes outside of the institution’s home country have a greater weight than domestic votes. Thus, the self-enhancement bias and modesty bias afflicting domestic evaluations of Eastern and Western institutions, respectively, may be balanced out by opposing biases from non-domestic respondents. Still, *in-group bias* may manifest in other ways beyond in-country bias among researchers from Western and non-Western institutions alike. For instance, academics may preferentially support colleagues who share their cultural, racial, or ethnic background—even when these individuals are part of diasporas outside their home country.

In survey research, such cultural divergences may result in *differential item functioning (DIF)*, where the same question (e.g., pertaining to reputation) may be understood in vastly different ways by respondents of differing demographic characteristics or cultural backgrounds (Brady, 1985; King et al., 2004). One method for correcting such bias would be the usage of *anchoring vignettes* where, in the context of a reputation survey, respondents would be asked to rate the reputation of one or more fictional universities based on descriptions of relevant characteristics such as research productivity, learning environment, etc. Answers to this “anchor” can then be used as a correcting mechanism to adjust subsequent evaluations of real universities.

Respondents may lack awareness of the specific institutions and academic standards in countries outside of their own, making it challenging for them to accurately assess quality in unfamiliar contexts. Even more concerning is the potential for stereotypes – both positive and negative – about the quality of institutions and the graduates they produce. For example, respondents from Western institutions may perceive universities in less developed countries (LDCs) as uniformly lower quality compared to those

in wealthier nations. This bias is often noted in studies of African universities, which have been penalized in global rankings despite substantial progress and accomplishments due to perceptions tied to historical and economic factors (e.g., Cloete, 2014). Studies suggest that respondents may unconsciously apply these stereotypes, underestimating the quality of institutions in unfamiliar regions or those perceived as less academically developed (Altbach, 2015; Hazelkorn, 2015). Here, again, the usage of anchoring vignettes or other corrective mechanisms for differential item functioning (where differences pertain not to reputation as a whole, but to universities in a particular region of the world) may mitigate bias.

3.2. Halo Effect

Established prestigious universities may benefit from a halo effect, where existing reputations are perpetuated regardless of current performance (Fidler and Parsons, 2008; Marginson, 2007). This effect can skew rankings and misrepresent the actual quality of institutions. Similarly, the perception of one positive trait (e.g., prestige) may lead to the assumption that the institution excels in other areas as well (Thorndike, 1920). This effect may exist for many aspects of the university experience, including in the context of teaching reputation.

In the context of a reputation survey, this may manifest as a form of *recency bias* where respondents rely heavily on past information (in this case, a university's perceived reputation in the prior year) when making judgments. This may also manifest as *anchoring bias* if panel respondents simply default ('anchor') to their evaluation of a university in the previous year. To mitigate this, methodologies such as *conjoint analysis* can be employed (Rao, 2014; Hainmueller et al., 2014). This particular survey design would allow researchers to isolate the specific attributes that drive respondent preferences, reducing the influence of historical or generalized reputational factors by focusing on concrete, comparable characteristics across institutions. Other methods exist for eliciting the *revealed preferences* of college students for specific universities: Avery et al. (2014), for example, demonstrate how to use a tournament-style experiment to aggregate students' preferences for universities they have been admitted to into a preference-based ranking of colleges.

3.3. Subjectivity

Reputation is inherently subjective because it is based on personal perceptions, which can vary widely depending on individual psychological dispositions, social status, race, and other demographic factors. *Social identity theory* posits that people are more likely to view institutions affiliated with their social group more favorably, leading to differential reputational evaluations based on factors such as class and race (Tajfel and Turner, 1986). Similarly, *confirmation bias*—the tendency for individuals to favor information that aligns with their pre-existing beliefs—can result in established institutions being perceived more positively regardless of their current performance (Nickerson, 1998).

On the institutional side of the equation, research highlights how institutions can actively shape their reputations through strategic branding and media exposure, influencing public perceptions independent of actual performance (Goffman, 1959; Schlenker, 1980). Institutions with more resources can invest

heavily in *reputation management*, skewing the public's perceptions and undermining the objectivity of reputation-based metrics in university rankings (Fombrun and Shanley, 1990).

These forms of subjectivity make it difficult to establish objective criteria for what constitutes a 'good' reputation, as the concept relies heavily on intangible factors such as prestige, recognition, and familiarity. In QS's own trade publication, reputation is conceded to be an "indefinable metric" that "depends a lot on brand identity, word-of-mouth, and other factors, such as how sustainable the university is perceived to be" (Gilmore, 2024). For these very reasons, the measurement of reputation in university rankings can be fraught.

3.4. Incentives

Personal incentives such as professional relationships, collaborations, or potential future benefits can also influence how respondents evaluate institutions. For example, an academic might rate an institution higher if they have ties with its faculty or administrators, in the hope of maintaining positive relationships or future opportunities. *Reciprocity bias* occurs when individuals rate an institution favorably because they expect a similar favor in return, further muddying the objectivity of the evaluation process. These biases may even affect the earlier sampling stages of reputation surveys: while THE develops the sampling frame (i.e., list of survey invitees) themselves, QS asks institutions to submit 400 names per year to build the list. Here, institutions can strategically curate their nomination lists to include individuals who are more likely to respond and provide favorable feedback. For instance, they may choose collaborators who perceive the institution as a prestigious partner or include names from countries that are underrepresented compared to the United States and Europe, given that votes from these regions carry higher weight in the overall evaluation.

Although QS does rely heavily on universities for academic and employer contacts for participation in reputation surveys, it is commendable that some ranking systems, like QS, do not allow self-nominations or exclude them from reputation surveys. This helps reduce the bias that might come from institutions attempting to inflate their own standing. However, even with such exclusions, there may still be incentives for reciprocal nominations or favoritism towards one's alma mater. Respondents may be inclined to support institutions with which they have personal or professional connections, which can distort the reliability of reputation scores, even if the institution is not their own.

Lastly, *cross-validating* reputation survey results with external data could help identify and correct discrepancies, leading to more reliable assessments. To our knowledge, none of these features have been applied to the three systems' survey methodologies; given the benefits demonstrated in previous research, we believe their inclusion would be valuable.

Reputation is ultimately downstream of more substantively meaningful constructs related to quality. In fact, these factors shape reputation itself: according to QS's own international student survey in 2023, the majority of respondents considered a high graduate employment rate to be the primary driver of a university's reputation. This raises the question of why these upstream factors, which have intrinsic value to students, are not given greater emphasis over reputation – a concept not without its merits, but complicated by the challenges of subjective perception and measurement.

4. Usage of Bibliometric Data

In addition to reputation, international ranking systems place significant conceptual emphasis on research productivity, primarily through the usage of bibliometric data.

Bibliometric indicators, such as the number of publications, citations, or the h-index, offer several advantages in assessing research output (Hirsch, 2005). First, they provide quantifiable metrics, allowing for consistent comparisons across institutions and disciplines (Moed, 2005). Bibliometric data can also highlight the impact of a researcher or institution by reflecting the extent to which their work is recognized and utilized by peers, serving as a proxy for academic influence and knowledge dissemination (Bornmann and Leydesdorff, 2014). Moreover, these datasets are often accessible via specialized vendors such as Elsevier's Scopus (or open source as in the case of OpenAlex) and can be collected over time, allowing for longitudinal analyses of research performance. The availability of large bibliometric databases like Scopus or Web of Science ensures broad coverage of academic fields, further enhancing their utility in global comparisons. This makes bibliometric data particularly useful for institutions and policymakers aiming to track progress or allocate resources based on measurable research outcomes.

Despite their advantages, the use of citation-based measures presents several challenges. As we detail below, much work has been done to address these biases, however there are no magic bullet solutions.

4.1. Statistical Biases

One common issue is the overrepresentation of certain fields, as research-intensive subjects like the natural sciences and medicine tend to produce more papers and receive more citations than fields like the humanities (Moed, 2005; Cronin and Sugimoto, 2014), often unaccounted for in even popular indices like the h-index (Hirsch, 2005). This creates an uneven playing field where universities strong in these areas appear more impactful than those excelling in less citation-heavy fields. Another bias arises from the geographic concentration of high-impact research in wealthier nations, which can skew rankings to favor institutions from these countries (Cronin and Sugimoto, 2014).

One way to address these biases is through *normalization by subject or field*, ensuring that citation counts are adjusted based on the typical citation patterns in each academic discipline. Other adjustments include *normalization by university size, longevity, location, or type*, accounting for the different scales, 'performance periods,' and specializations of institutions. QS and THE rankings normalize their bibliometric measures by subject or field, though none of the three rankings in question adjust by longevity.[4] ARWU, on the other hand, employs a separate index to capture the impact of social science publications, though there is little detail on whether and how disparate indices across fields can be compared.

Biases can also be introduced through *self-citations* or *institutional collaborations*, where researchers repeatedly cite their own work or that of close collaborators, inflating citation counts without necessarily

[4] For the 2026 ranking cycle, QS notes "that [they] have refined our normalization methodology. This update was introduced to achieve a more precise assessment of institutions at the indicator level" (Quacquarelli Symonds, 2025). It is possible that these refinements incorporate adjustments to better account for such differences across institutions.

reflecting broader research impact. Purging self-citations from bibliometric measures or mitigating their weight relative to other citations is one way to mitigate this issue. The QS ranking, for instance, excludes self-citations from its citation-based metrics.

4.2. Need for Multiple Measures

When using bibliometric data to assess research productivity, it is essential to recognize that no single metric can fully capture the multifaceted nature of academic research output. As Barari et al. (2024) summarize, the usage of proxy variables in particular calls for the combination of multiple measures in order to most precisely model the underlying target construct (e.g., research productivity).

Here, the Times Higher Education (THE) ranking system is effective at incorporating multiple dimensions of research productivity and influence, providing a more nuanced and balanced evaluation of institutions (Appendix Table A4). For instance, THE's "Citation Impact" metric assesses research influence by tracking the average number of times a university's published work is cited globally, using Elsevier's Scopus database. Importantly, the data are normalized to account for variations in citation volume across different subject areas and blends both country-adjusted and non-country-adjusted citation scores, offering a more equitable evaluation across geographic contexts. THE's "Research Strength" metric focuses on the Field-Weighted Citation Impact (FWCI) but limits the influence of outliers – papers with exceptionally high citation counts. By capturing the 75th percentile of a university's FWCI, this metric offers a more representative picture of an institution's research quality. Finally, the "Research Productivity" metric, which measures the total papers published normalized by staff size, subject, and institution size, allows for the evaluation of not just the volume but also the efficiency of research production, giving insight into how well an institution translates research resources into high-quality outputs.

In contrast, other ranking systems such as the ARWU and QS rankings employ more limited bibliometric indicators, which may fail to capture the full scope of research productivity. ARWU, for example, relies heavily on metrics like the number of highly cited researchers, papers published in Nature and Science, and the total number of papers indexed in the Science Citation Index-Expanded and Social Science Citation Index. While these metrics emphasize high-impact research, they overlook other aspects of productivity, such as research output in fields not typically associated with high citation counts or publication in highly prestigious journals. Similarly, QS focuses on the total citations received per faculty member over a five-year period, which can disproportionately favor larger institutions or those active in fields with high citation activity, without offering the same level of normalization seen in THE's approach. This singular focus on citations also fails to account for how efficiently institutions convert resources into publications or how diverse their research output might be across different disciplines.

The differences between these ranking systems highlight the need for multiple measures to fully capture multiple dimensions of research productivity. A combination of metrics that account for influence, quality, and volume – like THE's citation impact, research strength, and productivity indicators – provides a more comprehensive assessment of an institution's academic output. Ranking systems that rely too heavily on one aspect, such as citations alone, may miss important subtleties and fail to offer a fair representation of the complex landscape of global research.

4.3. Need for Robust Indices

To circumvent the aforementioned statistical biases (and others) as well as to capture further dimensions of research productivity, *indices*—higher order measures composed of multiple individual measures such as citation counts—could be used.

A prominent example is the h-index (Hirsch, 2005), though it is not currently used by any of the major ranking systems. The h-index, as it was originally proposed, has useful properties, such as jointly capturing researcher productivity and citation impact. However, it lacks crucial features like field-specific normalization, making it necessary to apply external adjustments or use blended indices (e.g., as with ARWU, between the Social Science Citation Index and the Science Citation Index) to account for differences across disciplines. Additionally, the h-index does not differentiate between authors' individual contributions in multi-author papers, nor does it account for researchers at different stages of their careers (Egghe, 2006). Other limitations include its insensitivity to the citation context, as it doesn't distinguish between positive or negative citations (Bornmann and Daniel, 2008), and its vulnerability to artificially inflated citations through self-citations or citation "cartels" (Schreiber, 2007). These shortcomings highlight the need for more nuanced metrics when assessing research impact.

Another example is the *A-Index* (Cabrerizo et al., 2009) which assigns relative credits to different co-authors based on their contributions. This method promises a more nuanced view of research impact by considering the distribution of effort among co-authors, rather than assigning equal credit for all papers. Such measures could help mitigate the over-crediting of institutions where researchers frequently collaborate on multi-authored papers.

Other alternative indices include the *g-index* (Egghe, 2006), which enhances the traditional h-index by allocating more weight to highly cited papers. The g-index is calculated based on the cumulative citations of an individual's most cited papers, offering a better balance between quantity and quality by recognizing the impact of high-performing publications without neglecting lower-cited work. This metric is particularly useful for researchers or institutions whose most influential papers are significantly more impactful than the rest of their output.

The *m-index* (also referred to as the m-quotient), which adjusts the h-index based on the number of years a researcher has been active, is another option (Harzing, 2012). This index helps account for the career stage of researchers, mitigating the advantage that more senior researchers with longer publication histories may have over early-career researchers.

Relative to the h-index, these alternative indices can provide a more balanced and multifaceted view of research productivity by addressing issues such as author contributions, paper impact, and career stage. Incorporating such indices into university ranking systems would offer a more holistic and equitable assessment of research output, ensuring that institutions are evaluated on a broader set of performance dimensions.

4.4. Research Quality as a Proxy

The significant weight on research outputs and the relative insignificance or absence of teaching quality or learning may be due to the *availability* of comprehensive data rather than its relevance to students'

preferences. In other words, ranking providers may be prioritizing what *can* be measured, rather than what *should* be measured to holistically represent their stakeholders' priorities. This emphasis may inadvertently reward institutions with strong research programs, which may not necessarily correlate with the quality of education provided to students. In fact, evidence suggests that students may choose universities with strong research programs not explicitly for the research itself, but because research prestige serves as a proxy for other desirable factors such as teaching quality and economic outcomes (Bok, 1986; Siow, 1997).

Recent survey evidence in the U.S. further confirms that teaching is a significant aspect of the college experience to many college-goers. When asked what makes the "best" college or university in a 2023 U.S. survey fielded by The Chronicle, respondents were much more likely to answer, "it has professors who are excellent teachers," than citing high graduation rates or good-paying jobs (McMurtrie, 2023). Although rarely directly captured in international rankings (in the THE and QS rankings, it is captured through indirect, proxy measures such as instructor/student ratio and institutional income), this aspect of education could hypothetically be systematically observed through careful, institution-level data collection as Campbell (2023) shows. Over the course of ten years, Campbell and her team observed 732 instructors across nine U.S. schools, evaluating academic rigor and active learning approaches in their teaching – both known to improve student learning outcomes. Campbell found these markers more common among instructors at regional state universities than at flagship or private research institutions, which tend to score higher in traditional rankings like U.S. News & World Report (Zimmerman, 2024). Though this finding pertains to the American higher education context, the implication for international rankings is that highly ranked global institutions may similarly fall short on measures of teaching quality compared to universities otherwise penalized for low research productivity. Further research would be needed to verify this hypothesis, although executing the same study in the international context would be infeasible. Importantly, as we noted in Section 2, simply comparing raw measures of teaching quality would not suffice: given the vast differences in the lengths, styles, and goals of undergraduate instruction across countries, further methodological considerations would be required to retain comparability in such measures between institutions.

Bibliometric measures, when used, must be carefully adjusted for bias. It is more critical that the attention paid to bibliometrics be proportional to their conceptual importance, when other dimensions, such as teaching quality, better align with what students value in their educational experience. If, however, the purpose of international rankings is to monitor (and incentivize improvements to) the performance of universities along research and reputational dimensions, current practices are defensible but require that ranking systems be clear about the interpretations of rankings.

5. Conclusion and Recommendations

In this report, we applied the construct validity framework from Barari et al. (2024) to assess the validity of international college rankings. While many issues are shared across domestic and international systems, certain challenges are more pronounced in international rankings due to the inherent constraints of cross-national comparisons across a much wider and varied set of institutions than domestic rankings. As the number of students seeking education abroad increases, the importance of these rankings will

inevitably grow for both institutions and students. To address the shortcomings identified in our assessment, we propose the following reforms:

Clarify the purpose of ranking. If international rankings aim to comprehensively reflect student preferences, the heavy emphasis on prestige and research outputs may overshadow other critical factors, such as teaching quality and accessibility. While institutional reputation is undoubtedly a key concern for both domestic and international students, aspects like curriculum flexibility and employment outcomes also play a crucial role in student decision-making (Obst and Forster, 2022). Conversely, if the primary goal of these rankings is to serve a narrower function—either catering specifically to students prioritizing prestige and research intensity or primarily serving as a benchmarking tool for evaluating institutions as research centers—then the current methodological focus may be appropriate. If so, ranking organizations must clearly communicate these objectives. Without such transparency, users may misinterpret rankings, erroneously equating differences in institutions' research performance with differences in institutions' absolute value, leading to potentially misguided decisions.

Personalize rankings and include customization elements. Rankings are consumed by many different audiences, each with different preferences over the possible conceptual dimensions of educational quality. International and domestic students, for instance, may be given vastly different financial aid packages to the same institution, face different geographic constraints both during and after graduation, and, as such, may have entirely different criteria for evaluating possible colleges to attend. As such, rankings should offer personalization features, allowing users to filter results by region, discipline, or specify their particular goals.

Leverage good practices from exemplary domestic institutions. Country-specific higher-ed educational ranking systems, such as the UK League Tables, offer valuable features that can inform or even be incorporated into international rankings. Drawing on the success of these models can help make international rankings more reliable and contextually appropriate for students and stakeholders. Current practices from QS to verify data against IPEDS (U.S.), HESA (U.K.), and individual university websites should be replicated and documented with clarity and detail.

Improve the usage of international reputation surveys and bibliometrics. To address biases in cross-cultural comparisons, rankings should employ techniques such as differential item functioning in surveys, which are well-documented in international assessment research (King et al., 2004). With bibliometric data, normalizations by subject, region, or field must continue and be expanded, ensuring accurate comparisons across institutions worldwide.

Our previous recommendations continue to hold, which include: ***empirical validation of the rankings*** (noting the added barriers of collecting international benchmarks), ***establishment of quality and transparency standards*** (potentially proposed, implemented, and/or enforced by international NGOs or governing bodies), ***communicating when differences between institutions are meaningful*** (which may include the usage of uncertainty statements such as confidence intervals and/or assigning institutions into tiers rather than individual ranks).

References

Altbach, P. G. (2012). The Globalization of College and University Rankings. *Change: The Magazine of Higher Learning, 44*(1), 26–31.

Altbach, P. G. (2015). Academic Inbreeding: Local Challenge, Global Problem. *Asia Pacific Education Review, 16*(3), 317–330.

Avery, C., Glickman, M., Hoxby, C., & Metrick, A. (2013). A Revealed Preference Ranking of U.S. Colleges and Universities. *The Quarterly Journal of Economics, 128*(1), 425–467.

Barari, S., Newsom, E., Park, J. E., & Paddock, S. (2024). College Ranking Systems: A Methodological Review. *NORC Reports.* https://www.norc.org/content/dam/norc-org/pdf2024/college-rankings-review.pdf

Bekhradnia, B. (2016). International University Rankings: For Good or Ill? (Vol. 89). Oxford: Higher Education Policy Institute.

Bok, D. (1986). Toward Education of Quality. *Harvard Magazine,* 49–64.

Bornmann, L., & Daniel, H. D. (2008). What Do Citation Counts Measure? A Review of Studies on Citing Behavior. *Journal of Documentation, 64*(1), 45–80.

Bowen, W. G., & Tobin, E. M. (2015). *Locus of Authority: The Evolution of Faculty Roles in the Governance of Higher Education.* Princeton University Press.

Bradburn, N. M., Cartwright, N., & Fuller, J. (2017). A Theory of Measurement. In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation* (pp. 73–88).

Brady, H. E. (1985). The Perils of Survey Research: Interpersonally Incomparable Responses. *Political Methodology, 11*(June), 269–290.

Bornmann, L., & Leydesdorff, L. (2014). Scientometrics in a Changing Research Landscape. *Science and Public Policy, 41*(3), 458–462.

Cabrerizo, F. J., Herrera-Viedma, E., & López-Herrera, A. G. (2009). A New Index for the h-Index With an Application to the Ioannidis Dataset. *Scientometrics, 81*(2), 335–345.

Calderon, A. (2020). New Ranking Results Show How Some Are Gaming the System. *University World News.* https://www.universityworldnews.com/post.php?story=20200612104427336

Campbell, C. (2023). *Great College Teaching: Where It Happens and How to Foster It Everywhere.* Harvard Education Press.

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students. *Psychological Science, 6*(3), 170–175.

Cloete, N. (2014). The South African Higher Education System: Performance and Policy. *Studies in Higher Education, 39*(8), 1355–1368.

Cronin, B., & Sugimoto, C. R. (Eds.). (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact.* MIT Press.

Docampo, D. (2011). On Using the Shanghai Ranking to Assess the Research Performance of University Systems. *Scientometrics, 86*(1), 77–92.

Docampo, D. (2013). Reproducibility of the Shanghai Academic Ranking of World Universities Results. *Scientometrics, 94*(2), 567–587.

Docampo, D., & Cram, L. (2014). On the Internal Dynamics of the Shanghai Ranking. *Scientometrics, 98*(2), 1347–1366.

Docampo, D., Egret, D., & Cram, L. (2022). An Anatomy of the Academic Ranking of World Universities (Shanghai Ranking). *SN Social Sciences, 2*(8), 146.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Egghe, L. (2006). Theory and Practice of the g-Index. *Scientometrics, 69*(1), 131–152.

Fidler, B., & Parsons, C. B. (2008). World University Ranking Methodologies: Stability and Variability. *Higher Education Review, 40*(3), 15–34.

Florian, R. V. (2007). Irreproducibility of the Results of the Shanghai Academic Ranking of World Universities. *Scientometrics, 72*(1), 25–32.

Fombrun, C. (1996). *Reputation: Realizing Value From the Corporate Image.* Harvard Business School Press.

Fombrun, C. J., & Shanley, M. (1990). What's in a Name? Reputation Building and Corporate Strategy. *Academy of Management Journal, 33*(2), 233–258. https://doi.org/10.5465/256324

Goffman, E. (1959). *The Presentation of Self in Everyday Life.* Doubleday.

Gilmore, J. (2024). Talking About Reputation. *QS Insights Magazine, 1*(18), 11. https://magazine.qs.com/qs-insights-magazine-18/talking-about-reputation

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis, 22*(1), 1–30.

Harzing, A. W. (2012). Reflections on the h-Index. *Business & Leadership, 1*(9), 101–106.

Hazelkorn, E. (2015). *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence.* Springer.

Heine, S. J. (2001). Self as Cultural Product: An Examination of East Asian and North American Selves. *Journal of Personality, 69*(6), 881–906.

Heine, S. J., & Hamamura, T. (2007). In Search of East Asian Self-Enhancement. *Personality and Social Psychology Review, 11*(1), 4–27.

Hirsch, J. E. (2005). An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations.* Sage Publications.

Iacobucci, D. (2013). A Psychometric Assessment of the *Businessweek, U.S. News & World Report,* and *Financial Times* Rankings of Business Schools' MBA Programs. *Journal of Marketing Education, 35*(3), 204–219.

Institute of International Education. (2023). Open Doors 2023 Fast Facts. https://opendoorsdata.org/wp-content/uploads/2023/11/Open-Doors-2023_Fast-Facts.pdf

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review, 98*(1), 191–207.

Kuh, G. D., Kinzie, J., Schuh, J. H., & Whitt, E. J. (2011). *Student Success in College: Creating Conditions That Matter.* John Wiley & Sons.

Kosmulski, M. (2006). A New Hirsch-Type Index Saves Time and Works Equally Well as the h-Index and g-Index. *ISSI Newsletter, 2*(3), 4–6.

Light, R. J. (2001). *Making the Most of College: Students Speak Their Minds.* Harvard University Press.

Marginson, S. (2014). University Rankings and Social Science. *European Journal of Education, 49*(1), 45–59.

Markus, H. R., & Kitayama, S. (1991). Culture and the Self: Implications for Cognition, Emotion, and Motivation. *Psychological Review, 98*(2), 224–253.

McMurtrie, B. (2023). Americans Value Good Teaching. Do Colleges? The Evidence Doesn't Look Good. *The Chronicle of Higher Education, 70*(3), 24–33.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation.* Springer.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw-Hill.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Obst, D., & Forster, J. (2022). Perceptions of European Higher Education in Third Countries: Outcomes of a Study by the Academic Cooperation Association (ASA): USA. *Institute of International Education (IIE).* https://www.iie.org/wp-content/uploads/2022/12/International-Students-in-the-US.pdf

Quacquarelli Symonds. (2024). How to Enter the QS World University Rankings: Your Questions Answered. *QS.* https://www.qs.com/insights/articles/how-to-enter-qs-world-university-rankings/. Accessed February 1, 2025.

Quacquarelli Symonds. (2025). QS World University Rankings. *QS.* https://support.qs.com/hc/en-gb/articles/4405955370898-QS-World-University-Rankings. Accessed September 3, 2025.

Rao, V. R. (2014). *Applied Conjoint Analysis.* New York: Springer.

Ross, D. (2023). World University Rankings 2024: Changes to Our Methodology. *Times Higher Education.* https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2024-changes-our-methodology

Saaty, T. L. (1987). Rank Generation, Preservation, and Reversal in the Analytic Hierarchy Decision Process. *Decision Sciences, 18*(2), 157–177.

Schlenker, B. R. (1980). *Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations.* Brooks/Cole Publishing Company.

Schreiber, M. (2007). Self-Citation Corrections for the Hirsch Index. *Europhysics Letters, 78*(3), 30002.

ShanghaiRanking. (2024). ShanghaiRanking's Academic Ranking of World Universities 2024 Press Release. *ShanghaiRanking.* https://www.shanghairanking.com/news/arwu/2024. Accessed February 1, 2025.

Siow, A. (1997). Some Evidence on the Signalling Role of Research in Academia. *Economics Letters, 54*(3), 271–276.

Tajfel, H., & Turner, J. C. (1986). The Social Identity Theory of Intergroup Behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of Intergroup Relations* (pp. 7–24). Nelson-Hall.

Thorndike, E. L. (1920). A Constant Error in Psychological Ratings. *Journal of Applied Psychology, 4*(1), 25–29. https://doi.org/10.1037/h0071663

Times Higher Education. (2024). Methodology for Overall and Subject Rankings for the Times Higher Education World University Rankings 2025. https://www.timeshighereducation.com/sites/default/files/breaking_news_files/the_2025_world_university_rankings_methodology.pdf. Accessed February 1, 2025.

Triandis, H. C. (1995). *Individualism & Collectivism.* Westview Press.

Zimmerman, J. (2024). Opinion: College Rankings Are Leaving Out the Most Important Factor. *The Washington Post.* https://wapo.st/3NwFpSc

Appendix

This report evaluates the three international ranking systems – QS World University Rankings, Academic Ranking of World Universities (ARWU), and Times Higher Education (THE) World University Rankings – based on the most recently available news articles, methodology reports, and other public statements made by each provider, at the time of writing (reflecting the 2024-2025 ranking cycle). The sources (along with their timestamps) are provided in Table A0. *We note that prior to the publication of this report, both QS and THE published its rankings for the 2025-2026 cycle, which include updates to methodological decisions shown here including the exact measures considered in each category, category weights, and the processing of measures (e.g. normalization of institution-level indicators)*.

Table A0. Consulted Sources for Evaluation of International Ranking Systems (2024-2025)

Ranking System	Title	URL	Type of Source	Published / Last Updated
QS World University Rankings	"QS World University Rankings"	https://support.qs.com/hc/en-gb/articles/4405955370898-QS-World-University-Rankings (note: this contains many links to other pages detailing specific methodological components, categories, and/or subcategories)	Methodology report	October 2nd, 2024
	"Understanding the Methodology: QS World University Rankings" (Staff Writer)	https://www.topuniversities.com/university-rankings-articles/world-university-rankings/understanding-methodology-qs-world-university-rankings	Press release / blog post	August 2nd, 2024
Academic Ranking of World Universities (ARWU)	"ShanghaiRanking's Academic Ranking of World Universities Methodology 2024"	https://www.shanghairanking.com/methodology/arwu/2024	Methodology report	Unknown
	"ShanghaiRanking's Academic Ranking of World Universities 2024 Press Release"	https://www.shanghairanking.com/news/arwu/2024	Press release	August 15th, 2024

Times Higher Education (THE) World University Rankings	"Methodology for Overall and Subject Rankings for the *Times Higher Education* World University Rankings 2024"	https://www.timeshighereducation.com/sites/default/files/the_2024_world_university_rankings_methodology.pdf	Methodology report	September 2023
	"World University Rankings 2024: methodology" (Duncan Ross)	https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2024-methodology	Press release / blog post	September 20th, 2023
	"World University Rankings 2024: changes to our methodology" (Duncan Ross)	https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2024-changes-our-methodology	Press release / blog post	September 20th, 2023

Table A1. Universe of Institutions in Each Major International Ranking System (2024-2025)

Ranking System	Eligibility Criteria (candidacy for ranking)	Inclusion Criteria (in final ranking)	Exclusion Criteria (in final ranking)
QS World University Rankings	• **Institution Type**: Must be an autonomous university or higher education institution authorized to grant degrees in multiple disciplines at both undergraduate and postgraduate levels with at least three graduated classes. • **On Campus Instruction**: Deliver all or part of degree programs on campus. • **Broad Program Offerings**: Must teach at least two of the following broad faculty areas: • Arts & Humanities • Engineering & Technology • Life Sciences & Medicine • Natural Sciences • Social Sciences & Management • **Narrow Program Offerings**: Must offer at least in two narrow subjects in each broad faculty area (as per QS subject classification) with at least three graduated classes. • **Institutional Autonomy**: Must be autonomous and not a branch campus of another institution. • **Academic Staff**: Must employ a minimum number of full-time academic staff. • **Justification**: For new entrants (not eligible or included in previous QS rankings), the university is required to provide "an argument or statement of why the institution should be included based on an objective comparison between their university and other institutions (in their country/region) that are already included in the ranking."	• **Research Output**: Must have at least 100 papers published in Scopus-indexed journals over a five-year period. • **Reputation:** Must be in the top 20% of institutions globally for Academic Reputation measure. • **Conditionality for Small Institutions:** If an institution has fewer than 5000 students must be one of the following: • Be in the top 1000 globally for Academic Reputation, Employer Reputation, or Citations per Faculty • Be in the top 900 in two of the above • Be in the top 800 for one of the above • **Rank Truncation/Tiering:** Ranks in the 2024 QS ranking appear to be truncated at [1401+], [1201, 1400], [1001-1200], [951-1000], in increments of 50 until 800 which increments in 10's until 600 (there are no scores for universities in these bands, only 'N/A'); after that certain rankings are skipped while others are repeated (e.g., both University of Missouri, Columbia and Sunway University are ranked #586 but there is no #583 university), but scores are present.	• **Vocational Institutions**: Institutions offering only vocational diplomas or not awarding degrees are excluded. • **Branch Campuses**: Generally, branch campuses are not considered independent and are excluded.

Academic Ranking of World Universities (ARWU)	• **Research Recognition**: Universities that have Nobel Laureates, Fields Medalists, Highly Cited Researchers, or papers published in *Nature* or *Science*. • **Research Output**: Universities with a "significant" number of papers indexed by the Science Citation Index-Expanded (SCIE) and Social Science Citation Index (SSCI)	• **Rank Truncation**: Only the top 1,000 ranked universities are published.	• **Non-University Institutions**: Institutions that may meet the eligibility criteria that are not universities (e.g., colleges, vocational schools) are excluded.
Times Higher Education (THE) World University Rankings	• **Research Output**: Must have published at least 1,000 research publications over five years (2018-2022), with a minimum of 150 publications per year. • **Institution Type**: Must be a higher education institution that teaches undergraduates. • **Subject Breadth**: Must cover a range of subjects and publish in sufficient numbers across THE's 11 subject areas (e.g., not more than 80% of all output from an institution from one area). • **Institutional Cooperation**: Must submit data required for evaluation, including a valid response to THE's annual institutional questionnaire.	• **Data Validation**: Institutions must pass THE's data validation processes.	• **Non-Participation:** Institutions that have requested not to participate in the ranking are excluded. • **Critical Missing Data:** Must not have more than two subcategory measures empty, unavailable, or withheld. • **Critical Missed Eligibility Criteria:** If some eligibility thresholds are not met, university will be excluded from ranking, but listed as a "reporter". • **Custom Exclusion Rules**: Methodology notes that universities "must not be featured in the custom exclusions list" with no further elaboration.

Table A2. Structure of the QS World University Ranking System (2024-2025)

Category	Weight	Category Measure(s)	Adjustments	Explanation/Justification
Academic Reputation	30%	Averaged response(s) to five faculty-area questions from worldwide **QS Academic Survey.**	• **Familiarity Weights**: Nominations are weighted based on respondents' regional and faculty familiarity and adjusted by the year of response to ensure relevance. • **Regional and Country Weights**: Weights are applied to nominations to prevent overrepresentation of certain regions or countries, ensuring fair representation globally. • **Domestic Weighted Count**: Domestic nominations are weighted and adjusted according to the number of institutions and response volume within a country, reflecting competition levels. • **Normalization and Transformation**: Both domestic and international counts are normalized to scores out of 100, combined using specified weights, and transformed to minimize the impact of outliers. • **Year Weights**: Nominations from the past five years are utilized, with decreasing weights for older responses (5th year at 25%, 4th year at 50%, and the most recent three years at 100%).	"The indicator not only illuminates the quality of an institution's research, but also their approach to academic partnerships, their strategic impact, their educational innovativeness and the impact they have made on education and society at large." "The assumption [behind equal weighting] is that, in a typical international comprehensive university, each of these faculty areas represents a **roughly equitable share** of activity."
Employer Reputation	15%	Responses to the **QS Employer Survey** assessing institutions' graduate employability.	(Similar to above)	"The majority of undergraduate students leave university in search of employment after their first degree, making the reputation of their university amongst employers a crucial consideration." "We remain the **only** major ranking to focus on this vital aspect of a student's educational journey."
Faculty/Student Ratio	10%	Ratio of full-time equivalent academic staff to number of full-time equivalent students.	• **Exclusion:** Administrative and support staff are excluded from FTE staff totals. • **Substitution:** When full-time equivalent students are missing, total students are used.	"The more academic staff resources are made available to students, for teaching, supervision, curriculum development, and pastoral support, the better the learning experience should be."

Citations per Faculty	20%	Total citations received by all papers produced by an institution over a five-year period per faculty member.	• **Institutional normalization**: Citation count is divided by the number of individuals in the faculty in order to take into account different sizes of institution. • **Paper-type exclusions**: Exclude certain content types (defined by Elsevier Scopus) from our analysis. • **Self-citations exclusion**: Exclude citation of an author's own work by said author • **Faculty area weighting/normalization**: Equalize the influence of research in our five key faculty areas, so that each contributes 20% to the final indicator.	"The indicator is a reflection of the volume of citations being achieved on average by an institution's academic staff. A higher volume of citations suggests that academics at those institutions are publishing in respected journals, engaging in strong collaboration and working on topics that merit a wide readership."
International Faculty Ratio	5%	Proportion of faculty members who are international.	• **Inclusion/exclusion**: Includes staff from Mainland China for Hong Kong universities; offshore exchange students and distance learning students are excluded. • **Dual citizenship rule**: In the case of dual citizenship, the deciding criteria should be citizenship obtained through birth, or first passport obtained.	"An institution attracting a sizeable population of international academics sees benefits in terms of its research and teaching diversity and **collaborations**. In addition, if an institution is attracting a sizeable number of overseas staff, it suggests that it has a **positive reputation** and is viewed as a good place to work. Institutions with high numbers of international staff can also benefit from wider international research networks due to the **connections** that their international academics bring with them, so a high score in this indicator hints at an open and collaborative academic environment."
International Student Ratio	5%	Proportion of students who are international (defined as foreign nationals, based on citizenship, who spend at least three months at the institution).	(Same as above)	"If an institution is attracting a sizeable population of international students this has benefits in terms of networking, cultural exchanges, a more diverse learning experience, and alumni diversity. In addition, if an institution is attracting a sizeable number of overseas students it suggests that it has a **positive reputation** and is viewed as a good place to study. This can be reinforced if graduates return to their home country with a positive experience to relay to future prospective students."
International Research Network	5%	Diversity of an institution's international research collaborations, measured using the Margalef Index.	(In order of application) • **Log-scale normalization:** Count of international countries/territories is divided by the natural logarithm of the distinct count of international partners. • **Min-max normalization:** Values for each faculty are scaled using min-max normalization from 1-100.	"The indicator measures how **diverse and rich** an institution's research network is by looking at the number of different countries represented, and whether these relationships are renewed and repeated."

			• **Aggregation:** Scaled measures are averaged. • **Z-score normalization**: Final measures are z-scored. • **Min-max normalization**: Final z-scores are further rescaled (to an unknown scale).	
Employment Outcomes	5%	Combined measure of alumni impact (share of "impactful graduates" on a domestic and global level) and graduate employment (% of graduates who go on to paid (non-voluntary) full or part time work within 15 months of finishing degree).	• **Log-scale normalization:** Applied to Graduate Employment Index prior to "draw in outliers and to ensure that the Graduate Employment Index component does not unduly influence the final score when compared with Alumni Impact Index". • **Combination:** Measures are multiplied.	"For many students a successful career is the **primary goal** of their university education and therefore is important to measure an institution's track record in this field. Equally an institution who produces graduates who go on to achieve success in fields such as the arts, politics, business etc. can point to their role in the development of those careers."
Sustainability	5%	Combination of measures from (1) institutional disclosures, (2) reputation survey responses, and (3) third party sources including the UN, UNESCO and the World Bank.	• **Weighting:** Applied to three different subcategory measures, each a weighted combination of several constituent measures.	"[The indicator] evaluates the social and environmental impact of universities as not only centres of education and research, but also as **major employers**. Sustainability is an increasingly important issue for students when picking a study destination and QS is proud to be the **first major** university ranking provider to include it as an indicator in our core rankings."

Table A3. Structure of the ARWU Ranking System (2024-2025)

Category	Weight	Category Measures	Adjustments
Alumni	10%	Number of alumni winning Nobel Prizes and Fields Medals.	• **Min-max normalization:** highest-scoring institution is given a 100 score, others calculated as a percentage of the top score. • **Time weights:** "Different weights are set according to the periods of obtaining degrees…100% for alumni obtaining degrees after 2011, 90% for alumni obtaining degrees in 2001-2010, 80% for alumni obtaining degrees in 1991-2000, and so on, and finally 10% for alumni obtaining degrees in 1921-1930." • **Truncation:** "If a person obtains more than one degree from an institution, the institution is considered once only."
Award	20%	Number of staff winning Nobel Prizes in Physics, Chemistry, Medicine, and Economics and Fields Medals in Mathematics.	• **Time weights:** (see above) • **Proportional credit allocation:** "If a winner is affiliated with more than one institution, each institution is assigned the reciprocal of the number of institutions. For Nobel prizes, if a prize is shared by more than one person, weights are set for winners according to their proportion of the prize."
Highly Cited Researchers	20%	Number of Highly Cited Researchers selected by Clarivate Analytics.	N/A

Nature and Science Papers	20%	Number of papers published in *Nature* and *Science*.	• **Proportional credit allocation:** "To distinguish the order of author affiliation, a weight of 100% is assigned for corresponding author affiliation, 50% for first author affiliation (second author affiliation if the first author affiliation is the same as corresponding author affiliation), 25% for the next author affiliation, and 10% for other author affiliations." • **Corresponding author decision rule:** "When there are more than one corresponding author address, we consider the first corresponding author address as the corresponding author address and consider other corresponding author addresses as first author address, second author address etc. following the order of the author addresses."
Publications	20%	Total number of papers indexed in Science Citation Index-Expanded and Social Science Citation Index.	• **Special weights:** "When calculating the total number of papers of an institution, a special weight of two was introduced for papers indexed in Social Science Citation Index."
Per Capita Performance	10%	Combination of all other scores.	• **Weights**: Each of five indicators are weighted (unknown values). • **Substitution**: "If the number of academic staff for institutions of a country cannot be obtained, the average number of academic staff for world top 1000 universities will be used for all institutions in this country." • **Normalization**: Score is divided by the number of full-time equivalent academic staff.

Table A4. Structure of the THE Ranking System (2024-2025)

Category	Weight	Category Measure(s) (and subcategory weights)	Adjustments (note that all final measures are z-score normalized)	Explanation/Justification
Teaching (the learning environment)	29.5%	• Reputation survey (15%) • Staff-to-student ratio (4.5%) • Doctorate-to-bachelor's ratio (2.25%) • Doctorates awarded per academic staff (6%) • Institutional income per staff (1.75%)	• **Log-scale normalization**: "Only non-zero values will be standardized using a logarithmic function" • **Missing value imputation**: "universities that received no [reputation] votes are scored a zero for this metric."	• **Doctorate/staff ratio:** "As well as giving a sense of how committed an institution is to nurturing the next generation of academics, a high proportion of postgraduate research students also suggests the provision of teaching at the highest level that is thus attractive to graduates and effective at developing them." • **Institutional income**: "This measure of income indicates an institution's general status and gives a broad sense of the infrastructure and facilities available to students and staff."

Category	Weight	Category Measure(s) (and subcategory weights)	Adjustments (note that all final measures are z-score normalized)	Explanation/Justification
Research (volume, income and reputation)	29%	• Reputation survey (18%) • Research income (6%) • Research productivity (5%)	• **Reputation weights**: "Each year is calculated as the number of global research votes from respondents of the reputation survey, weighted by subject and country to be representative of the distribution of academics globally." • **Log-scale normalization**: "Only non-zero values will be standardised using a logarithmic function" • **Missing value imputation**: "universities that received no [reputation] votes are scored a zero for this metric." • **Income adjustments**: "Research income is weighted by subject, adjusted for PPP, by the total subject weighted number of academic staff, and is normalized after calculation." • **Productivity proportional credit allocation (with threshold):** To "give credit for cross-subject research that results in papers being published in subjects where a university has no staff … we will reassign the papers to subjects where there are staff. We will do this proportionally according to the number of staff in populated subjects, and according to the median publications per staff for populated subjects. We will have a maximum threshold of the proportion of papers that we are willing to reassign (10% of the total of papers)."	• **Research income**: "This is a somewhat controversial indicator because it can be influenced by national policy and economic circumstances. Income is crucial to the development of world-class research, and because much of it is subject to competition and judged by peer review, our experts suggested that it was a valid measure."
Citations (research influence)	30%	• Citation impact	• **Normalization**: impact normalized by field and blended with country-adjusted measures of citation count.	"Our research influence metric looks at universities' role in spreading new knowledge and ideas." "[Normalization] means that institutions with high levels of research activity in subjects with traditionally high citation counts do not gain an unfair advantage."
International Outlook	7.5%	• Proportion of international students (2.5%) • Proportion of international staff (2.5%) • International collaboration (2.5%)	• **Country Normalization**: all measures normalized to account for the country population's size. • **Subject Weighting:** Co-authorship metric "generated by dividing the total subject weighted number of publications with at least one international co-author by the total subjected weighted number of publications."	N/A

Category	Weight	Category Measure(s) (and subcategory weights)	Adjustments (note that all final measures are z-score normalized)	Explanation/Justification
Industry Income (knowledge transfer)	4%	• Industry income per staff member (2%) • Patents (2%)	• **Patent Field Normalization:** The patent "measure is subject weighted to avoid penalizing universities producing research in fields low in patents." • **Patent Staff Normalization:** "We also normalize [patents] by the sum of academic and research staff."	"An institution's ability to help industry with innovations, inventions and consultancy has become a core mission of the contemporary global academy. This category suggests the extent to which businesses are willing to pay for research and an institution's ability to attract funding in the commercial marketplace – useful indicators of institutional quality." "[The patents] metric recognizes the extent to which universities are supporting their national economies through technology transfer. It measures the count of patents citing an entity's published research."