NORC at the University of Chicago

FINAL REPORT
September 2024

College Ranking Systems:
A Methodological Review

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

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

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

Table of Contents

About NORC **3**

Executive Summary **3**

1. Introduction **6**
1.1. Illustrative College Ranking Systems 7
1.2. Components of College Ranking Systems 8

2. Background: Construct Validity **9**

3. Review: Conceptualization **11**
3.1. Differing Definitions of "College" 11
3.2. Ambiguous Characterizations of "Ranking" 13
3.3. Lack of Content Validity 14
3.4. Lack of Criterion Validity 15
3.5. Lack of Measurement Invariance 16
3.6. Lack of Scale Reliability 16
3.7. Formative vs. Reflective Constructs 17

4. Review: Data **18**
4.1. Data Quality Issues 19
4.1.1. Proxy Measures 19
4.1.2. Missing Data 19
4.1.3. Data Lag 20
4.1.4. Total Survey Error 21
4.1.5. Simple Aggregation 21
4.1.6. Lack of Documentation 22
4.2. Review of Data Sources 22
4.2.1. Integrated Postsecondary Education Data System (IPEDS) 22
4.2.2. College Scorecard 23
4.2.3. Surveys 24
4.2.4. Third-Party Vendors 26
4.2.5. Other Public Sources 26
4.2.6. Direct Disclosure from Institutions 27
4.3. Review of Specific Measures 27
4.3.1. Graduation Rates 27
4.3.2. Student Debt 28
4.3.3. Value-Added Earnings 30
4.3.4. Fellowships 31
4.3.5. Financial Resources 32

5. Review: Methodologies **33**
5.1. The Equal Intervals Problem 33
5.2. Subjective Weights 34

5.3. Categorization and Transformation 35
5.4. Uncertainty 37

6. Conclusions and Recommendations 38
6.1. Empirically validate existing rankings 39
6.2. Establish robust quality and transparency standards 40
6.3. Consider ratings as an alternative to rankings 40
6.4. Improve the visual presentation of the final measure 41
6.5. Consider personalizing college rankings 41
6.6. Develop a reform agenda with stakeholders at the table 42

Funding Statement 42

References 42

Appendix 46
Appendix 1. Additional Details about Five Exemplar College Ranking Systems 46
Appendix 2: Total Survey Error Framework 53
A.2.1. Validity 53
A.2.2. Measurement Error 54
A.2.3. Processing Error 54
A.2.4. Coverage Error 55
A.2.5. Sampling Error 55
A.2.6. Nonresponse Error 56
A.2.7. Adjustment Error 57
Appendix 3. Federal Committee on Statistical Methodology (FCSM) Data Quality Framework 57
Appendix 4. Berlin Principles on Ranking Higher Education Institutions 59
A.4.1. Purposes and Goals of Rankings 59
A.4.2. Design and Weighting of Indicators 60
A.4.3. Collection and Processing of Data 61
A.4.4. Presentation of Ranking Results 61

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

Higher education has long been viewed as a stepping-stone for individuals to achieve professional and personal goals. To demystify the options within the higher education landscape, prospective students and their families often turn to college rankings. This is especially true in an era of increasingly expensive college costs (Franke, 2017). Several major college rankings providers have recently revised their methodology to better reflect these cost concerns, along with making other updates. While this effort may be laudable, any proposal for substantive changes in rankings raises the question of the underlying qualities encapsulated by the rankings and whether the ascribed methodology can accurately capture those qualities.

Given these potential issues, we conducted a methodological review of five prominent college rankings systems: U.S. News & World Report (USNWR), Wall Street Journal (WSJ), Forbes' Top College list (Forbes), The Times Higher Education World University Ranking (THE), and the QS World University Ranking (QS).

Our overarching approach is to assess the construct validity of college rankings. 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 includes a review of conceptual, data, and methodological factors within this framework that are central to developing a methodologically sound ranking of colleges, drawing examples from our five illustrative ranking systems. Here, we summarize five key issues spanning the conceptual, data, and methodological aspects of construct validity along with an initial recommendation to address each (expanded further in Section 6).

1. There is not one clear or stable set of concepts underlying the college ranking systems. In the absence of a universal definition of what college rankings measure, each provider defines its own target concept and updates it periodically. Still others rely on entirely socially determined inputs for their ranking scores – for example, public opinion surveys of college-going students – implying that rankings are undergirded by a perpetually moving target rather than a stable, identifiable concept or set of concepts. Without a stable and clearly defined conceptualization – which may encompass multiple constituent concepts – it is difficult to understand which individual measures map onto which relevant conceptual dimensions and how or whether to combine them into a single measure. More broadly, as we explain in Section 3, without clarity on a measure's underlying conceptualization, it is difficult or impossible to draw complete or coherent comparisons about the real-world institutions it is aiming to describe. Finally, as we discuss in Section 6.2, it is entirely permissible with social scientific measurement to allow for concepts *to be defined by* the measures used to assess them, rather than the concepts *to define* the measures.

However, to evaluate construct validity effectively, it is crucial to distinguish which of the two practices is being employed.

We recommend, among other things:

a) reconceptualization of ranking systems, clarifying the basket of well-identified concepts encapsulated in each ranking and whether they define or are defined by the choice of measures,
b) disaggregation of ranking systems altogether that produce multiple measures where each maps onto narrower, but more clearly defined and interpretable concepts, and/or
c) thorough assessment of the validity of the existing measures through typical psychometric procedures (e.g., criterion validity, predictive validity, and others discussed in Section 3).

2. Issues of data quality limit the ability to capture even well-defined concepts. Even if the underlying concepts are well-defined psychometrically and substantively, the ability to capture them accurately relies on the credibility and quality of the many data sources that feed into college ranking systems. College ranking systems often draw from self-reported administrative data or from surveys administered by third parties that often lack transparency because they may exercise broad discretion in the public release of methodological information. Colleges themselves are also known to "game" the ranking system, selectively disclosing data, or reactively allocating funds and programming to achieve greater scores in constituent ranking categories (Thaddeus, 2022; Mendoza, 2022). In contrast, data from other public (particularly government) sources usually must comply with specific methodological and disclosure requirements, thus generally deemed to provide high-quality estimates of their intended populations. Even so, such high-quality data is suited to its intended original purpose; its usage as a placeholder (or 'proxy') for other populations (e.g., imputing the characteristics of all students based on high-quality estimates of those characteristics for federal loan recipients) may undermine the overall quality of measurement.

Given the variety of data sources used in college rankings, each with its own measurement challenges, we outline several relevant data quality considerations a rankings developer or consumer should consider (Section 4.1). We apply these considerations to six key measures in our illustrative systems: graduation rates, student debt, value-added earnings, fellowships, and financial resources (Section 4.3). We show that college ranking systems often use data that do not fully align with what is being measured or indirectly proxy for more informative, but unavailable or inaccessible, data for the task at hand.

As a result, we recommend:

a) greater transparency about these possible data quality issues, and additionally,
b) implementation of processes to improve data quality, where possible, including institutional audit mechanisms, more rigorous statistical correction procedures, and a review of data sources according to the quality frameworks we discuss throughout the report.

3. Many methodological procedures applied to the data involve subjective decisions. Assuming conceptual and data issues are resolved, there are concerns pertaining to the operationalization of measures — that is, the specific procedures and transformations applied in the process of constructing ranks. At many of these decision points, there is a wide degree of discretion available to the ranking constructors: the definition of an eligible collegiate institution for ranking, the classification of majors into STEM vs. Non-STEM, the substitution of certain measures for others (e.g., graduation rates for standardized test scores as done by USNWR), and the truncation of colleges below a certain threshold, to name a few (discussed further in Section 3). Perhaps the most consequential methodological decision that is subjectively

derived (in nearly all of our exemplars) is the weight assigned to each substantive category of measures (e.g., financial resources vs. reputation).

While *changes* to these weights over time and their *relative* magnitudes are often justified, there is little rationale for the precise percentages assigned to each category. It is not clear that such precise weight values are based on measurement considerations.

We recognize that it is impossible to remove subjective decision-making from the ranking methodologies altogether. As such, we recommend that:

a) decisions are better justified from a measurement perspective,
b) sensitivity of rankings to changes in such decisions are determined and ideally featured as an interactive tool, and/or
c) external validation of the resulting measures guide how these decisions are made.

4. There is insufficient characterization of the uncertainty in each numeric ranking. As with any statistical measure – a weather forecast, an IQ test, or a nation's GDP – there is uncertainty about the precise 'true' location of each measured unit. By our count, none of the major ranking providers communicate this uncertainty, potentially misleading consumers about the degree of differentiation between institutions. For example, it is unclear whether differences between colleges #1 through #3 (or between college #1 and #30) are large enough to be statistically significantly different and substantively meaningful. Relatedly, rankings fundamentally face an issue we have dubbed the 'equal interval problem' – an implication that the distance between adjacent ranks is uniform. Namely, rankings create the illusion that colleges ranked #1 and #2 are *equally* distinct in the underlying rankings concepts as colleges ranked #29 and #30, for example. Beyond uncertainty around the final product, there is uncertainty associated with nearly every measure used in the construction of each rank (e.g., uncertainty resulting from low response rates in a student evaluation survey).

We discuss a number of these sources of uncertainty (Section 5.4 and elsewhere) at various stages of measurement and recommend that the providers quantify these sources, including in the ranking itself, and visually display this information in public reports. If the uncertainty is too great to make meaningful comparisons among the rankings, we recommend that providers either truncate ranks into tiers (as they already do in some cases) in a way that conveys meaningful differences or consider abandoning ranks altogether in favor of categorical ratings or other alternative systems.

In summary, many aspects of college rankings suffer from a lack of construct validity. This starts with the conceptualization of what is measured, and then extends to *how* concepts are measured in data elements drawn from various data sources, and finally to how the data/information are processed and presented using various methodologies. Improving college ranking systems will require addressing all these areas. We believe there is a need for using data to explore these limitations, develop communication strategies at the appropriate level to inform the public of these issues, and examine approaches that might improve college rankings.

1. Introduction

Higher education has long been viewed as a stepping-stone for individuals to achieve professional or personal goals. In an era where higher education is becoming increasingly expensive, prospective students and their families often turn to college rankings to demystify the higher education landscape and options (Franke, 2017). A survey from Art & Science Group, a higher education consultancy, showed that "58 percent of college-bound high school seniors considered rankings to some degree" (Blinder, 2024). College rankings can significantly influence student perceptions and, consequently, their application decisions, with higher-ranked institutions often seeing an increase in applications and thereby enhancing their selectivity and perceived prestige (Bastedo & Bowman, 2010). Prospective students often interpret rankings as a measure of educational quality, believing that attending a higher-ranked institution will result in better educational and career outcomes. This belief can lead students to prioritize ranked institutions in their application process, potentially at the expense of other factors that may be more relevant to their personal and educational needs (McDonough et al., 1998). Indeed, studies have demonstrated a correlation between changes in an institution's ranking and corresponding shifts in the quantity and quality of its applicant pool, underscoring the impact of rankings on student choices (Ehrenberg, 2003).

At the institutional level, the effects of college rankings can be both beneficial and challenging. Ideally, rankings could serve as a tool for self-evaluation and improvement, as they often include indicators related to teaching quality, research output, and student satisfaction. Institutions use these metrics to identify areas of weakness and allocate resources accordingly (Sauder & Espeland, 2009). However, unintended consequences include some institutions adopting strategies focused on improving their position in rankings at the expense of broader educational goals. For example, some colleges have been revealed to manipulate data or focus disproportionately on metrics that are rewarded by ranking methodologies, such as selectivity, faculty resources, or alumni giving rates, even if these metrics do not necessarily translate to educational quality or a positive impact on student outcomes (Sauder & Espeland, 2009). This behavior, often referred to as "gaming the rankings," raises concerns about the validity of college rankings.

In this report, we review conceptual, data, and methodological factors that challenge the methodological rigor of college rankings. Chief among them is the question of construct validity, namely *what exactly are the college rankings providers trying to measure and how accurately and completely are they able to measure it?* Whether it has to do with classical notions of educational quality, market-oriented concepts like customer satisfaction or value generation, or bigger normative considerations like equity, mobility, and redistribution, clarity is needed on the measurement goals of college ranking systems. Without a universal definition of what college rankings are trying to measure and an agreed-upon measurement framework, the college ranking systems each use their own subjective evaluations to derive the overall score used in the rankings. As we show in Section 3, this poses a threat to construct validity, the scientific gold standard for the usage of any numerical scale that aims to describe abstract qualities like 'market value' or 'educational quality.'

Secondly, the credibility and quality of multiple data sources that feed into college ranking systems are often questionable. College ranking systems often draw from self-reported administrative data or from surveys administered by third parties that often lack transparency because they may exercise broad discretion in the public release of methodological information. In contrast, data from other public (particularly government) sources usually must comply with specific methodological and disclosure

requirements and are deemed as of high quality. Nonetheless, we show that college ranking systems often use data elements that do not fully align with what is being measured (starting in Section 4.1) or indirectly proxy for more informative, but unavailable or inaccessible data for the task at hand (in Section 4.1.1). In Section 4.2, we provide an overview of commonly used data sources followed by a discussion of measurement properties of each in Section 4.3. For further reading, in Appendix 3, we introduce the Federal Committee on Statistical Methodology (FCSM) Data Quality Framework as a tool to evaluate data quality (one of several discussed in the appendix).

Finally, even when conceptual and data issues are resolved, there are concerns pertaining to the operationalization of measures — that is, the specific procedures and transformations applied in the process of constructing ranks (the focus of Section 5). Chief among them is that rankings, as opposed to ratings, imply that the distance between adjacent ranks is uniform, a problem dubbed the 'equal intervals problem' (discussed in Section 5.1). We further note that many methodological decisions that go into the operationalization of measures may be unfit for use in ranking systems.

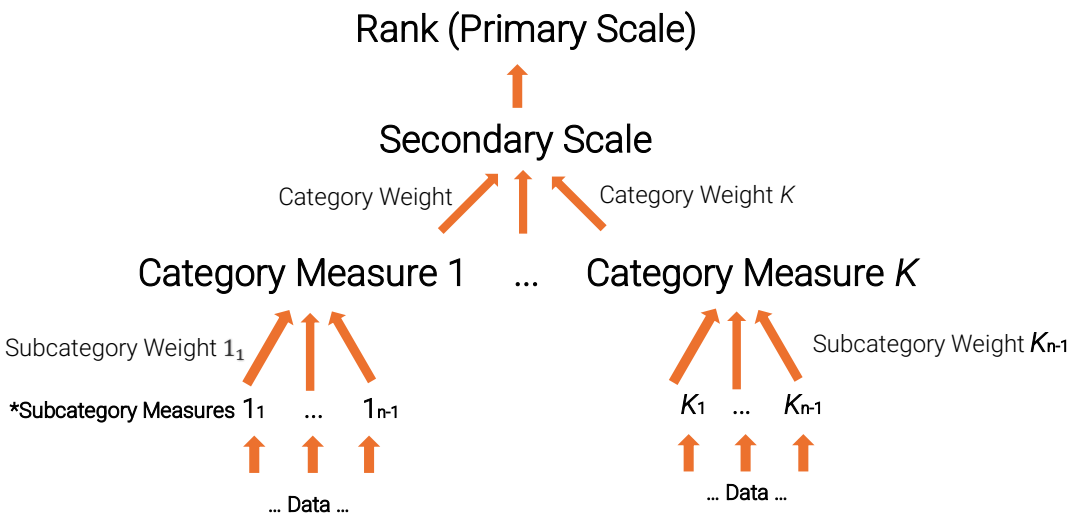
In sum, this report reviews a number of frameworks and tools to critically evaluate the concepts, data, and methodologies used in college ranking systems. Five college ranking systems were chosen as exemplars to illustrate a range of high-profile rankings. We will focus on evaluating a subset of measures and the data included in these systems and provide methodological recommendations and directions for future research.

1.1. Illustrative College Ranking Systems

In our report, we focus on five illustrative ranking systems: three domestic —The U.S. News & World Report (USNWR), Wall Street Journal (WSJ), Forbes' Top Colleges (Forbes) — and two international — the Times Higher Education (THE), and the QS World University Rankings (QS). In this section, we describe the structure of these systems.

Our evaluations of each of the five illustrative systems throughout this report are based on the public disclosures of methodology by the systems themselves. Links to these reports can be found in the footnotes of Appendix Table 1.

1.2. Components of College Ranking Systems



*In some ranking systems (e.g., USNWR), one or more category measures are composed of exactly one measure or data source (rather than multiple subcategory measures).

Figure 1. Graphical Illustration of College Ranking Structure

Definitionally, the primary quantity of interest in each ranking system is an ordinal scale for educational institutions denoting their integer **rank** (e.g., 1, 2, 3,...) according to some interpretation such as 'best national university' or 'best value' (more on this conceptualization in Section 3) as used by USNWR. We hereby refer to this as the **primary scale**, although we note that it may be interchangeably called a 'rank.'

Each primary scale or rank is directly derived from some **secondary scale,** a numeric quantity that is sometimes but not always displayed alongside the rank. As Figure 1 shows, this secondary scale is created from the aggregation of multiple measures spanning different categories (each hereby referred to as **category measures**) of institutional characteristics such as financial aid or instructional quality.[1] When the secondary scale is displayed to the end consumer (e.g., WSJ), it is typically normalized to some range (0-100) and referred to as a 'score'. Appendix Table 2 shows which components of the ranking system are displayed to the end consumer across the five illustrative ranking systems.

A crucial feature of the secondary score is that it is constructed via the application of a **category weight** (ranging from 0-100%) to each category measure. Appendix Table 3 summarizes the publicly available information about the category weights in each ranking system. **Some systems justified** *changes* to weights, but not the precise weight value, while others did neither. In four of our illustrative ranking systems, weights are based on subjective categories rather than data-driven values based on judgments of 'low' or 'high' values. The exception is the QS World Universities ranking system where weights are determined via surveyed student priorities. However, no details about how this information is used to construct the weight values could be found in QS methodology reports. A recent trend (USNWR 2024, WSJ 2024, and Forbes 2024) has been to increase weights from previous rounds allocated to student

[1] In the psychometric literature, what we have dubbed the 'secondary scale' may also be called a *composite measure*, meaning an aggregation of many different measures from possibly disparate data sources. We return to the literature on composite measures (and their implications for construct validity) in Section 3.7.

'outcome' categories (e.g., earnings) relative to educational experience or other 'input' variables (e.g., selectivity).

Finally, each secondary scale may be derived from multiple measures relevant to that category (also called 'indicators' or 'metrics' or 'indexes' by some systems, but hereby referred to as **subcategory measures**) also combined via weights (**subcategory weights**). For example, the Forbes 2024 Top Colleges list has one category score for 'return on investment' and one category score for 'academic success' and the latter is an aggregation of two subcategory measures: the number of recent graduates who have won prestigious academic fellowships and the number of recent alumni who have gone on to earn doctorate degrees. Some ranking systems use a handful of coarse, thematic categories (e.g., Forbes and WSJ), while in other cases (e.g., USNWR), there is only one variable or data source feeding into each category score, in which case there are no subcategory scores.

Each subcategory measure itself is drawn from **data**, defined here as quantified observations of the real-world: in the former case, the lists of graduates and alumni compiled by the institutions. Data are transformed into measures through a variety of methods ranging from simple counts (e.g., total number of award-winning alumni) to averages (e.g., average earnings) to more complex statistical models (e.g., estimates of value-added adjusting for regional and demographic characteristics relevant to each institution).

Having defined the visible components of a college ranking system, we next introduce the social scientific framework of construct validity to help us describe the underlying purpose of each measure in the system and evaluate whether this purpose is met.

2. Background: Construct Validity

What do college rankings measure? Indeed, before any specific datasets or statistical techniques can be brought to bear in the construction of numerical college rankings – or any measure for that matter – this question needs answering. Namely, the underlying *construct* that the measure is attempting to capture requires definition.

A **construct** may be tied to one or more *concept*s, which Bhattacherjee (2012) defines as "generalizable properties or characteristics associated with objects, events, or people," for example, a "person's attitude towards immigrants [or] a firm's capacity for innovation." While some concepts and the measures used to describe them may be precise and self-evident (e.g., cost), others may be broad, abstract, and even qualitative in nature (e.g., prestige) as well as not directly observable in the real world (e.g., economic value), though it may have implications for real-world phenomenon (e.g., earnings for graduates, consumer demand).

A construct differs from a concept in that it is specifically selected (equivalently, 'defined' or 'conceptualized') to explain a social phenomenon and, therefore, may be more case-specific or contextual in nature. Constructs also may encompass many different related but more general concepts. For example, a construct of usage in our present case of college rankings may be 'value to enrolled students' (rather than 'value' to some broader cohort) which may encompass concepts such as 'prestige' (or 'social value'), 'return on investment' (or 'economic value'). In the case of a multi-faceted construct, we

may dub these concepts *dimensions* of the construct. These concepts, in turn, have their own describable *features* or observable 'real world' phenomena. For example, the aforementioned dimension 'economic value,' one dimension of 'student value,' may have the associated feature of 'high salaries for student graduates.'

This definition of construct maps directly onto the structure of measures composing college rankings as illustrated in Figure 1. The precise mapping is as follows: within the structure of college rankings, the *construct* of interest maps onto the ranking or primary scale while the concept *dimensions* directly correspond to the category measures used to create the secondary scale and, finally, *features* of those dimensions map directly onto data used to construct the subcategory measures feeding into the category measures (or, in the case where there are no subcategories, the category measures directly). The analogous figure for Figure 1 is shown below, directly resembling the structure of college ranking measures in 'conceptual' rather than 'empirical' terms.

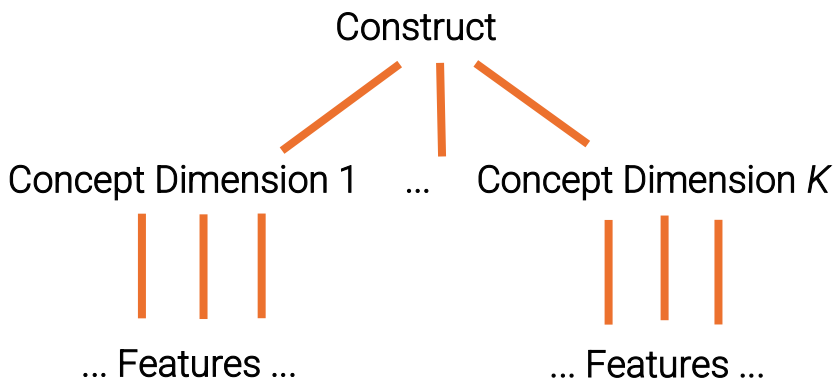


Figure 2. Conceptual Illustration of College Ranking Structure

Nevertheless, in order to make accurate and generalizable conclusions, evaluations, or comparisons about the construct of interest (e.g., college X is more valuable to enrolled students than college Y) that are not simply self-evident facts (e.g., college X costs more than college Y), the associated measures (e.g., reputation perceptions, statistically modeled salary estimates) used to draw those inferences must clearly, unambiguously and holistically capture different dimensions of the construct (e.g., social value, economic value). Moreover, the basket of concepts that make up a multi-dimensional construct should also be well-defined and clearly relate to each other and the overarching construct definition.

If this can be said, the concept and its measurement achieve the property of *construct validity* (Cronbach and Meehl, 1955). Otherwise, not only is our conceptualization invalid, but our conclusions are likely to be incomplete, inaccurate, or incoherent.

Though many frameworks and procedures for establishing construct validity exist, we draw from Bradburn, Cartwright and Fuller (2017) in the present context of college rankings. To demonstrate that a measure and its designated construct meet the standard of construct validity, Bradburn et al. (2017) require the following steps:

- Characterization: define the construct, including defining the conceptual dimensions that together comprise 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., student value for a single college) based on observable real-world features (equivalently, 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 then requires explaining how these requirements are met and harmonized: "the representation of the quantity or quality measured must be appropriate to the central features taken to characterize it; equally, the procedures adopted to carry out the measurement must be appropriate to the formal representation adopted, and we should have good reason to expect that within acceptable bounds of accuracy the values assigned indicate the values we aim towards" (Bradburn, Cartwright and Fuller, 2017). A valid construct that meets these three requirements has implications for real-world outcomes.

We identify several threats to construct validity across each of our five major ranking systems. Some of them are associated with choices of datasets or operations that result in omissions of relevant dimensions while others are simply failures to explain or clarify the features themselves. Deeper yet are threats to validity due to differing or unclear definitions of what a college itself is and the concepts that each ranking aim to characterize.

We discuss issues and threats to construct validity organized under the domains that draw from Bradburn et al. (2017): conceptualization (i.e., characterization), data (i.e., representation), and methodologies (i.e., operationalization). We first describe issues related to characterization and representation specifically that pose significant threats to the validity of college ranking systems in Section 3. We continue our discussion of representation and data quality issues as they pertain to individual data sources in Section 4 and home in on issues at the operationalization stage (methodologies) in Section 5.

3. Review: Conceptualization

We begin our critical assessment of the construct validity of college ranking systems at the conceptual level (or characterization in the Bradburn *et al.* framework). That is, we ask: Is there an identifiable construct that a ranking is meant to describe? Is it described fully? Are the boundaries with other related, but distinct constructs clarified? These questions must be answered with consideration of the specific measurement processes chosen for such description (or representation in the Bradburn et al. framework). There are at least six major issues in the conceptualization of college ranking systems, which we turn to now.

3.1. Differing Definitions of "College"

It is impossible to characterize the construct underlying a 'college ranking' without first establishing the concept of a 'college.' From the onset, each of the five systems defines 'college' itself differently. While most rankings rely on an external taxonomy commonly used in the educational research field to identify

these criteria (e.g., one of the Carnegie classifications), others, particularly international collegiate rankings (e.g., THE, QS), do not point to an external population frame, or listing of all institutions eligible for inclusion as predetermined using well-defined criteria. A summary of the eligibility and exclusion criteria used across the five illustrative systems is provided in Appendix Table 1.

In general, these criteria do not appear to be especially controversial; each set of inclusion rules forms a defensible definition of 'college' by most standards in educational research. On cursory inspection, the set of resulting colleges (particularly in the top ranks of each system) nearly perfectly overlaps. However, the possibility of different acceptable definitions of 'college' raises two issues in the conceptualization of 'college rankings.'

Firstly, the inclusion criteria for 'college' used by providers are often seemingly arbitrary thresholds that vary widely across ranking systems. Forbes, for example, only includes colleges with more than (exactly) 300 undergraduate students in its universe; other systems do not apply such a threshold. This conceptualization raises several questions. How many colleges that met all other criteria were removed simply because they had 299 students instead of 300? What happens if the floor is raised to 400 instead of 300? Not least of all is the question: why 300? One may infer that the choice of 300 (or its general ballpark) is motivated by 'measurement' concerns – that is, the tools to measure the dimensions of the construct-of-interest break down for colleges below a certain size. In the language of construct validity, the choice of measure (representation and operationalization) informs the boundaries of the construct (characterization). This is permissible by most construct validity frameworks. However, when definitions are based on strict cut-offs, care must be taken to (i) explicitly justify the choice of cut-off based on measurement or other considerations or (ii) demonstrate that measures and their substantive conclusions are not highly sensitive to the exact choice of cut-off (i.e., confirm that rankings are minimally changed by slight perturbations to student body threshold).

Second, inclusion criteria for 'college' are not just informed by general measurement concerns, but sometimes defined by the exact inputs used to construct ranking measures. For example, in the Times Higher Ed ranking, a college may only be considered for evaluation if its faculty publication measures – likely meant to capture the conceptual dimension of instructional quality – meet some minimal subject-wise thresholds as well as some minimal level of completeness in reporting. These productivity measures, however, are also used in the construction of the THE ranking scale. This potentially results in a circular conceptualization: a college must report a minimal and complete level of whatever it is that ranks a college highly.

As with the measurement of any social phenomenon, reasonable differences may exist about the exact contours of the underlying construct. Explicit rationale for the contours separating an institution that is a 'college' from an institution that is not, however, are seldom given by the five rankings, weakening their characterization.

Finally, flexible definitions of eligibility may result in changes in relative ranks based on nothing more than the inclusion or exclusion of certain institutions. In a given year, a college may rise in the ranks merely because institutions previously ranked higher were disqualified according to previously established criteria (e.g., a decline in enrollment falling below the existing threshold for student body size) or with the addition of new criteria (e.g., an added threshold for student body size). In addition to a conceptualization issue, eligibility criteria are thus among the many instances of methodological discretion that pose threats to construct validity which we explore further in Section 5.

3.2. Ambiguous Characterizations of "Ranking"

In general, there is considerable ambiguity in the construct underlying each of our exemplars' rank measures as well as their associated concepts. Providers describe what their rankings aim to capture in broad and usually relativistic terms – colleges high on their list, or on the list altogether, are the 'best' or 'top' colleges in America (or the world in the case of the international providers). The 2023 WSJ findings report begins by asking "what makes a college [...] great." The 2023 *Forbes* ranking similarly notes that its list showcases 500 of the "finest U.S. colleges."

We believe it is unlikely that such abstract qualities are the characterization goal of college rankings. While 'greatness' or being among the 'finest' may be recognizable *concepts*, this differs from a scientific *construct* in that it is neither specific to the context of study nor does it serve an explanatory purpose. Thus, such vague and comparative characterization begs the question: 'best,' 'top,' 'great,' or 'fine' in what?

Most of our five providers do, in fact, offer an answer to this question in their methodological reports, describing the related conceptual *dimensions* of constructs:

- The 2024 USNWR ranking system's underlying construct can be best interpreted as "[ability to offer the] top reasons that students attend...college or university" which include "academic reputation, cost of attending, and return on investment."

- The 2024 THE ranking system characterizes the underlying construct as an assessment of 'university performance on the global stage' and qualifies that their "rankings cover the three main areas of university activity: research, impact and teaching".

- The 2024 Forbes ranking system's underlying construct can be best interpreted as "[ability to] deliver on the promise of a quality education," of which includes "data on student success, return on investment, and alumni influence."

- The 2024 WSJ ranking system scores are based on student outcomes, the learning environment, and diversity.

- The 2024 QS ranking system does not explicitly describe its underlying construct but notes that it reflects the "focus of different stakeholders".

It should be emphasized that identifying the construct in each case, if even possible, requires reading into specific phrases or statements. Nowhere in any of the five methodology reports is there an explicit statement along the lines of 'our ranking system aims to capture [construct X]'. Arguably, in many instances, the basket of related dimensions at play (e.g., research, impact, and teaching) when, taken together, may or may not definitionally cover the purportedly over-arching construct (e.g., performance). Still, as per the requirements of the characterization step, features associated with the construct and its underlying conceptual dimensions are described by each provider.

To their credit, some systems define constructs by highlighting which concept dimensions and associated features do *not* belong to their construct-of-interest. For example, the headline of WSJ's methodology report itself reads 'Measuring Outcomes, Not Inputs,' which summarizes their elimination of "assumption(s) that the quality of education is largely dictated by how expensive it is to produce" and

"a greater emphasis on measuring the value added by colleges—not simply measuring their students' success, but focusing on the contribution the college makes to that success." Still, in the next section, we raise the possibility that these features may not completely 'cover' the features associated with the construct of interest.

Concerningly, some of the rankings appear to be *re-characterizing* previously established constructs, suggesting that the construct itself may not be fixed. WSJ notes that the 2024 rankings "no longer reward colleges' wealth or reputation in and of themselves;" in lieu, the rankings place "greater emphasis on measuring the value added by colleges." Other rankings note that changes in their approach apply to their methodology rather than the construct: The 2024 QS ranking methodology report boasts the largest ever "methodological enhancement" which reflect "different stakeholder's shift(s) in response to wider trends in education and society." The QS and USNWR rankings, notably, appear to conceptualize their rankings in terms of consumer demand. Hence it can be argued that the construct is inherently characterizing a 'moving target,' shifting with public opinion and therefore possibly requiring different representational choices from year to year. However, none of the ranking systems – QS and USNWR included – explicitly state this assumption nor have they done so in previous iterations of rankings, casting doubt on whether changes in datasets or procedures reflect (a) improvements in capturing a fixed construct or (b) adjustments to capture different constructs over time.

Providers often use evaluative language such as 'blame', 'reward', 'punish', and 'credit' – e.g., WSJ writes that they "no longer reward colleges' wealth or reputation in and of themselves". As Appendix Table A1 shows, there are numerous other examples of perceived fairness considerations cited in methodologies of the five exemplars including product differentiation (between the ranking systems themselves). Indeed, the ranking of colleges is conducted by competing firms in a marketplace where reputation, fairness, and product differentiation may be important considerations. Moreover, fairness can be an important consideration in measurement contexts such as educational assessment, where concerns about biases against certain historically 'under-assessed' subgroups may arise (Dorans and Cook, 2016). However, without a clear definition of the underlying construct being measured there is a great deal of ambiguity about the intent of these methodological decisions – improving the quality of the measurement or improving outcomes for individual institutions (or the providers themselves).

In summary, our assessment is that the five major college ranking systems adequately *describe* features of the ranking construct, but do not precisely or explicitly define it, leaving ambiguities in characterization. Even in the former case, these features may not be exhaustively captured by each ranking's representation and operationalization (we turn to this in Section 3.3). In our five example systems, the lack of a coherent, well-defined construct suggests that, alternatively, the individual category measures may map onto distinct concept dimensions that do not define one overarching construct. If so, it is unclear how to interpret a weighted combination of the measures corresponding to these concepts, particularly when they are combined in a highly subjective way (see Section 5.2). Moreover, many providers' admission of re-characterization over time raises the question of whether the construct is entirely socially determined, whether it is adjusted to positively reflect a handful of known institutions, or whether there is a single stable construct at all.

3.3. Lack of Content Validity

A measure achieves *content validity* (Cronbach and Meehl, 1955) when all conceptual dimensions of the construct, and the noteworthy features belonging to those dimensions, are fully and accurately represented by the measure and any construct-irrelevant contents are excluded. Content validity can be evaluated through the scrutiny and approval of experts who can identify and recognize the lower-level

features and higher-level dimensions (as shown in Figure 2) undergirding the construct-of-interest (Furr, 2011). In the case of our college ranking providers, because the construct itself is not adequately defined it cannot be determined whether the component measures of each ranking fully cover these dimensions or if, conveniently, the dimensions of the construct itself are narrowly defined by measures that are readily available.

Regardless of the true directionality, there are likely omissions of potentially relevant dimensions of the collegiate experience since ranking systems differ from *each other* in their choice of measures. Surveys of academic reputation and student perceptions are used in some ranking systems but not others: the USNWR and THE rankings explicitly omit any measures of positive or healthy student campus experience. Meanwhile, Forbes relies heavily on alumni performance using a unique set of achievement measures (described further in Section 4.2.4) not considered (or available) to other providers. The usage of academic reputation surveys fluctuates in usage ("gone is the survey of academics on schools' reputations," writes WSJ).

It is, thus, unclear whether these differences in the representational and operational stages are due to differing dimensions of concepts being captured by the ranking (e.g., value-added vs. perceived reputation), different subjective interpretations of the same construct (e.g., 'greatness' interpreted in terms of social prestige vs. economic mobility), differing subjective judgments of how to operationalize a fixed concept dimensions (e.g., using faculty output vs. faculty perceptions to measure academic quality), or other non-methodological considerations (e.g., concessions to favor certain stakeholders over others).

3.4. Lack of Criterion Validity

As Bradburn et al. (2017) also note, the measure used to represent a construct should be validated against other known measures that are substantively linked to the same construct. This is often called *predictive validity* when the scores are in a different domain and measured in the future or *concurrent validity* when the scores are measured contemporaneously in the same or a closely related domain (Cronbach and Meehl, 1955). These two types of validity, together, make up *criterion validity*. Kim and Shim (2019) assess criterion validity of liberal arts colleges rankings by examining their association with exposure of students to good educational practices.

Theoretically, it is argued that institutions such as universities exhibit path dependence, or a slow trajectory of change in uptake of new technologies, organizational practices, curricular reforms, and management ideas (Krücken, 2003; Schreyögg and Sydow, 2009). In essence, barring substantial events like insolvency or closures, we should expect that more things stay the same at any given university between consecutive years than change. Hence, it may be a reasonable assumption that numeric rankings of the same institution should correlate across consecutive years within the same system if they are able to holistically represent the university. The available literature on the validity of college university rankings includes analyses where this stability assumption is invoked (e.g., Tancredi et al, 2013; Iacobucci, 2013).

Demonstrating high levels of correlation or intercoder reliability (Cohen, 1960) within a particular year but across the systems could establish concurrent validity, assuming the central construct underpinning each system is similar. Setting aside the fact that the construct may not be the same, similar, or even identifiable across systems, this, too, is complicated due to differing definitions of an educational institution itself as we previously established. At present, we see no evidence that either form of criterion validity is considered in the construction of any of our five college rankings.

We did not find documentation from measure developers of how or whether criterion validity was a factor in the development of rankings. However, there has been some evaluation of criterion validity and other psychometric properties of published rankings in the academic literature, particularly for professional programs. Iacobucci (2013) examined aspects of MBA program rankings, including correlational analysis of program rankings over time, distinctions and overlaps among three high-profile sets of rankings with respect to what is being measured, and associations of rankings with outcomes such as employment. As important and valuable as these analyses are, Iacobucci (2013) notes some limitations, such as when outcome measures fluctuate due to broader trends (e.g., economic conditions influencing employment prospects).

3.5. Lack of Measurement Invariance

Another important property of an instrument specifically used for making comparisons between different groups of individuals, institutions, or repeated measures is *measurement invariance*, which establishes that the selected measure (or "representation" in the Bradburn et al. framework) has the same validity (e.g., criterion validity or content validity) across different groups for which the construct applies (Meredith, 1993). In our case, the groups in question may include institutions of varying sizes, institutions across regions, or institutions with varying degrees of first-generation students. Establishing measurement invariance would require demonstrating, for example, that the conceptual dimensions of small institutions are represented as accurately and fully (i.e., content validity) as the dimensions of large institutions.

A common framework for establishing measurement invariance for a selected measure involves structural equation modeling (SEM), or the estimation of unobserved properties of both individual units (e.g., Princeton University, Hillsdale College) and groups of units (e.g., National universities, Liberal Arts colleges) in a dataset. SEM could entail conducting a series of confirmatory factor analyses (CFA) to test if the same goodness-of-fit from a given structured model is achieved across all groups. In other words, this analysis would entail demonstrating that neither a unit's unobserved structural parameters nor the corresponding model's ability to predict observable outcomes for that unit is dependent on the group membership of that unit. At present, we find little to no evidence that measurement invariance was formally considered in the development of new measures, such as the ones derived from student/faculty/reputation surveys (discussed further in Section 4.2.3).

3.6. Lack of Scale Reliability

There are also concerns regarding the *scale reliability* of each ranking system, or the extent to which it produces consistent results if the procedure (collection of data, transformation into intermediate measures, and combination into overarching measures per the structure shown in Figure 1) is repeated (Nunnally, 1975), an important implication of construct validity.

Some researchers have explored stability of rankings over time to assess scale reliability to the extent possible, given that each college ranking scale is a static, yearly snapshot so it necessarily cannot be replicated in the following year. For example, Tancredi et al. (2013) examined stability and spread in primary care medical school rankings, concluding the considerable spread in scores over time for non-top 20 programs is larger than expected given the pace of actual changes occurring in medical training, calling into question scale reliability. Drawing from the test-retest correlations used in education, Iacobucci (2013) evaluated the reliabilities of MBA rankings by examining correlations over time, but

noting the analytic limitation that the students, recruiters, and other parties included in the rankings may be different individuals.

Assessing reliability in terms of the stability of ranks over time requires an assumption of stability in higher education and preferences widely held about higher education. While stability is a reasonable assumption, it becomes more challenging to assume stability when the input criteria for ranks changes over time and when there are important changes in the higher education landscape, such as recent changes to major rankings giving more weight to socioeconomic concerns of students. This problem is best described as *methodological drift*: Across the numerous, sometimes drastic changes to ranking methodologies each year, it is unclear whether the underlying construct, characterization, and/or representation, or simply the methodological procedure to accurately capture the same construct, is changing. A famous saying in survey research puts this problem in simpler terms: 'if you want to measure change, don't change the measure.' Notes from the UNSWR and the QS World University Rankings heavily imply that their target construct is consumers' (i.e., incoming students') preferences and values about collegiate experience and outcomes, which inherently change year-to-year. Other ranking systems (e.g., Forbes) neither concede that their target construct may be inherently dynamic nor commit to a fixed conceptualization of college quality.

Relatedly, college ranking is an inherently *thermostatic* process: institutions are responsive to their rankings and may attempt to omit, manipulate ('game'), or otherwise 'self-correct' the same information if asked to disclose it in a hypothetical repeat of the ranking process. Given the feedback mechanisms between stakeholders and the ranking providers, institutions may even lobby at various points of a ranking audit if their ranking was already made public. This, too, is a barrier to replicating college rankings both within and across years.

3.7. Formative vs. Reflective Constructs

A composite measure is constructed by combining several measures from possibly different types of data and is used as the basis for ranking in college ranking systems. When dealing with composite measures, it is essential to specify whether the underlying construct is formative or reflective, as this distinction impacts both the interpretation of the construct and the validity assessment methods used (Bollen & Lennox, 1991).

One pathway is to conceptualize "quality" captured in college ranking systems as a *reflective construct*, which exists independently of the specific measures that are combined into an overarching composite measure. This also means that the measures can be added or dropped without changing the meaning of the construct (Shwartz et al., 2015). This justifies the usage of indirect measures of the concept (as we further discuss in Section 4). Reflective constructs, however, require that the multiple measures used for its characterization be highly statistically correlated and conceptually related. In other words, the input measures of 'quality' must all point in the same direction to provide an unambiguous portrait of the underlying construct. The less this is true, the greater the uncertainty in the resulting primary scale and the lower the interpretability of comparisons between colleges along the primary scale.

Alternatively, "quality" in college ranking systems can be conceptualized as a *formative construc*t, which is formed *from* the measures (rather than one that *forms* the measures), which means the construct will change with any changes in the measures associated with the construct. Formative constructs are much more flexible in their interpretation. Instead of finding downstream observable implications of an unobservable construct (as in the case of reflective constructs), the analyst can simply

define the construct in terms of observable implications. For instance, the approach taken by WSJ in their college rankings appears to employ a formative construct. According to their methodology report, they “now put greater emphasis on measuring the value added by colleges” that, along with other variables such as “success in absolute terms” compromises student outcomes (weighted at 70%) which “rebalances” with other criteria in the composite score (learning environment at 20% and diversity at 10%).This language reflects a formative construct, as it defines "quality" based on the specific measures of student outcomes, the college's added value, and other measures of learning environment and student diversity directly shaping the construct around these indicators.

In simpler terms, a reconceptualization along these lines requires that we ask: Are student employment outcomes and perceived instructional quality great because the college is ‘great’ (reflective) or is a college ‘great’ because student outcomes and instructional quality are great (formative)? Current practice among the college ranking systems, similar to WSJ, resemble that of formative constructs rather than reflective constructs. However, beyond the need for simply ‘more transparency,’ it is crucial to clearly distinguish which of the two conceptualization approaches is being used for practical reasons of ensuring validity. Firstly, the type of construct has implications for the types of validity tests that need to be conducted. Bollen and Lennox (1991) illustrate that with formative constructs, conventional tests such as confirmatory factor analysis do not straightforwardly apply, as these constructs do not assume that indicators will covary. This necessitates alternative methods for assessing validity, such as examining the construct’s predictive validity or its ability to explain variations in outcomes of interest. Secondly, there are critiques within the psychometrics literature about the use of formative constructs altogether. Diamantopoulos and Winklhofer (2001) point out that formative constructs can be problematic because they rely heavily on the researcher’s choice of indicators, making them highly susceptible to biases and instability over time. Edwards and Bagozzi (2000) also argue that the interpretability of formative constructs is challenging because the relationships among indicators are not based on a shared latent trait, but rather on the chosen components that define the construct. This ambiguity can lead to difficulties in drawing consistent and reliable conclusions across different studies and contexts. The validity of formative constructs should be assessed to ensure that the chosen indicators accurately capture the intended dimensions of the overall construct.

To summarize the issues of conceptualization in college rankings: the central concept underlying each of the five college systems lacks clarity, making it difficult to assess the validity or stability of the corresponding representation schemes. It is indeed possible that there are multiple, potentially divergent concepts of interest. If so, it is unlikely that their corresponding measures can be straightforwardly combined into a single, interpretable scale. We discuss possible solutions to improve the conceptualization of college rankings (in either scenario) in Section 6.

4. Review: Data

To capture salient features of ranking systems requires evaluating the suitability of the data used in rankings for achieving the stated purposes of the systems. In this section, we provide exactly this assessment regarding the *representational* requirement of construct validity: “define the measure, or the process of assigning a number to each instantiation of the based on observable real-world features (data), that appropriately and fully describes the construct.”

We begin in Section 4.1 with six broad themes of data quality issues (4.1.1—4.1.6) that hamper many different data sources' abilities to produce measures that 'appropriately and fully' describe a reasonable construct of college rankings across all systems. In Section 4.2, we then delve into six commonly used data sources and demonstrate how these six issues (and others) apply to each provider in further detail. Following that, in Section 4.3, we zoom in even further to five specific measures produced from (and often across multiple) data sources, flagging how these six issues may distort the measurement properties of each.

4.1. Data Quality Issues

4.1.1. Proxy Measures

Nearly all category and subcategory measures shown in Figure 1 are *proxy measures* (also called *indicator variables*), meaning that they are observed, indirect measure of a concept when direct measures are not possible or available. Measuring *student intelligence* (unobserved) from observed *test scores* (proxy measure) is one of the oldest and most well-established examples (Spearman, 1904).

All proxy measures share one thing in common: They are imperfect representations of the underlying concept. This is a necessary trade-off in the analyses of institutions such as companies, universities, governments, and political parties which are complex, multi-faceted, but nevertheless clearly exhibit both quantitative and qualitative differences in their outputs (Adcock and Collier, 2001). Moreover, any one proxy measure may be inadequate in fully representing the unobserved concept, which poses a threat to content validity as described in the previous section. As a result, it is recommended that multiple proxy measures (rather than just one) are used to capture the concept of interest (Bollen, 2002).

While conceptually appropriate proxy measures enable us to describe the unobservable, they also demand a more complex operationalization to account for the increased uncertainty about the underlying concept, typically through the usage of statistical models (Bollen, 2002). We see little to no evidence these operational needs have been met in our systems as we discuss further in Section 5.4.

Appendix Table A4 provides a comprehensive (though not exhaustive) overview of the types of proxy measures utilized within categories of each ranking system. We discuss a few prominent examples of these in more detail in Section 4.3.

4.1.2. Missing Data

Missing data remains a persistent issue in data collection and measurement (Little and Rubin, 1989). Accordingly, a myriad of methods exists at the *operationalization* stage to handle missing values in variables of interest: substitution (e.g., via the sample mean, values from similar observations), deletion (e.g., row-wise or column-wise), or imputation (e.g., via a statistical model on observed values).

The accuracy of each method in predicting true values when they are unobserved depends on whether their corresponding assumptions about the *reasons* for missing-ness are met. Assumptions range from values missing completely at random (MCAR) – the most permissive though least realistic assumption – under which it is possible to analyze data only from institutions with complete data and not introduce missing data bias; to missing at random (MAR), under which an unbiased estimate can be obtained by

adjusting for missingness by using observed traits from other correlated measures; to not missing at random (NMAR), which is the most difficult-to-correct scenario, as it arises when the probability an observation is missing depends on its true but unobserved value and requires complex statistical modeling under particularly strong assumptions.

One example of missing data includes non-response to survey questions – for example questions around perceptions of the college environment asked of students in the College Pulse survey used by WSJ. Responses to this question are unlikely to be missing completely at random, as students with more negative perceptions or less strongly felt perceptions altogether are more likely to skip or refuse this question. Still, responses may be missing at random conditional on some observed traits such as academic performance or social satisfaction.

It is difficult to ascertain the quality of variables with missing values and their resulting measures without a transparent explanation of the underlying data-generating process – in the aforementioned example, the survey response mechanism – and an explicit assumption about the nature of missingness. On this criterion, the USNWR generally outperforms its peers, reporting that missing survey values from its internal statistical survey are handled via substitution from the equivalent variable in the IPEDS survey system. Further, the 2024 USNWR methodology report notes that if less than 50% of ACT/SAT scores are reported from an institution, the resulting measure of test performance is dropped altogether from the weighting – instead, the category weight for graduation rate is increased by 4% to fill the gap (on the basis that it is most strongly correlates with standardized test scores).

4.1.3. Data Lag

As mentioned in our discussions of scale reliability (i.e., reproducibility of rankings over time), the temporal nature of college rankings is a liability: 'changes in the measures' across consecutive years makes it difficult to 'measure change' in colleges between those years. The assemblage of college ranking measures also suffers from a different temporal issue, namely the mismatch in the time periods over which data is collected which we call *data lag*.

For instance, data released from official government institutions is often disclosed earlier than estimates from institutions themselves. Using one interchangeably for the other may result in an incoherent characterization of the institution in a given year. Taking the average of two different counts of a college's study body taken months apart may result in a number that never actually existed at any point during that period. These two endpoints may also differ *across* colleges. This might lead to a scenario in which two colleges are close in assigned ranks even though their underlying ranking data is collected over incomparable, non-overlapping time periods.

Moreover, when certain data are obtained during specific time windows during the academic year (e.g., the College Pulse survey is administered between January and May) the resulting measures may exhibit seasonal effects. In this example, student perceptions of their college may be more positive not due to institutional conditions but rather warming springtime weather.

There has also long been debate about the inherent stability of public opinion, particularly around evaluations of broad entities (Druckman and Leeper, 2012). In the case of evaluating ones' educational institution, students may be susceptible to a host of recency or availability biases based on the immediate positive or negative consideration they are able to draw on during a survey. We now turn to contextualizing this in a framework for assessing the data quality of survey data, given the use of surveys to collect data used in college rankings.

4.1.4. Total Survey Error

The most well-established framework of survey quality, *total survey error* (Groves, 2005) describes the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data. The Total Survey Error (TSE) framework focuses on the concept of total survey quality, where the key lies in the optimal allocation of resources to reduce the errors that may occur in surveys (Groves et al., 2009).

Appendix Section 2 provides a graphical overview of the Total Survey Error framework and the seven types of error canonically considered in survey practice. Here, we introduce and define a subset of errors for subsequent discussion in the next section:

- Measurement Error. Measurement error bias (also called systematic error) and measurement error variance (also called stochastic or random error) are two sources of measurement error in surveys. Measurement error bias can result from systematic deviation in responses from the true value due to poor measurement, such as misreading or misunderstanding a survey question. On the other hand, measurement error variance can result from variability across measurements.

- Processing Error. Processing error can also arise from the transfer of data from the recording into the analytic dataset, such as data entry errors in a disclosure form, misreading of respondent's writing, or rounding up/down of dollar figures.

- Coverage Error. Coverage error occurs when there is a gap between the target population and those represented in the sampling frame, which is the list of population members from which a sample is drawn. For example, if the sampling frame in the survey excludes specific individuals, such as those without cell phones, or students who live off-campus, the survey will suffer from coverage error.

- Sampling Error. Another source of error - closely related to coverage error - can arise in the process of sampling, which introduces variance in survey estimates. Sampling error can be reduced when a larger sample of the target population is taken.

The TSE framework emphasizes the importance of addressing these errors through a combination of optimizing data collection procedures and application of appropriate statistical methods.

4.1.5. Simple Aggregation

A common theme across multiple data providers and data elements is the aggregation of quantified observables into simple averages, totals, or medians. For example, average net tuition is used as a summary of the cost of attendance for a given university. This, however, masks another important property of such data elements: the spread or 'variance' of values observed in the world.

We illustrate the problem with aggregating away the variation with the example of tuition. College X's average tuition may be lower than College Y, rendering it a favorable category score for a conceptual dimension such as 'affordability.' Yet, College X's maximum tuition charged to students in the incoming cohort may be significantly higher than College Y. Alternatively, College X may have waived its tuition for more students than College Y which is counted in the average as "zero;" on the other hand, the average tuition for College X students who *pay any* tuition may be significantly higher than for College Y.

The "over-aggregation" of such data elements can, thus, mask inequities across groups represented in the measure. While a college may be performing well according to some summary of an indicator, there may be significant variation across student groups, departments, or programs. The incorporation of more disaggregate data into college rankings can better reflect the full distribution of such indicators, aiding in the overall construct validity of rankings.

4.1.6. Lack of Documentation

Finally, paralleling the lack of clarity at the conceptualization stage, we have broadly observed a lack of clarity in the exact origins, collection, processing, and assemblage of the datasets used across the ranking systems. This exacerbates the previous four issues since even efforts to correct for data quality issues cannot be evaluated if there is limited documentation on their execution.

Having identified five broad themes in data quality issues across our exemplar systems, we now zoom in on where they occur in the underlying data.

4.2. Review of Data Sources

We now highlight several key data sources used throughout the ranking systems that exemplify the data quality issues discussed above.

4.2.1. Integrated Postsecondary Education Data System (IPEDS)

IPEDS is a collection of 12 surveys administered to higher-educational institutions that are conducted annually by the National Center for Education Statistics (NCES) in the U.S. Department of Education. Widely considered the seminal dataset on higher-ed institutional characteristics in education research, IPEDS is the source of several categories of data used across the rankings:

- Institutional/human resources, e.g., institutional revenue and faculty salaries (used as proxies by USNWR),
- Enrollment, e.g., fall enrollment (used to construct retention rates by USNWR and Forbes),
- Admissions, e.g., median standardized test scores for each entering class (used by USNWR),
- Student outcomes, e.g., graduation rates (used by WSJ) and provisional graduation rates for Pell grantees (used by USNWR).

Other data elements from IPEDS may have been incorporated into the five systems; however, these are the usages most clearly disclosed across the five methodology reports.

One notable characteristic of the IPEDS data system is that it has a mandatory participation requirement for all institutions that receive Title IV federal assistance, thereby, can serve as a common data source for inclusion in the ranking systems. In theory, this should mitigate or altogether thwart the issue of missing data or coverage error (in the TSE framework). However, there is a small number of institutions, perhaps most famously Hillsdale College, that do not receive federal funds and are thus nowhere to be

seen in the IPEDS system. These institutions are typically small, religious liberal arts colleges which may categorically exempt from the rankings' population of interest (e.g., institutions of a specific Carnegie classification excluding religious colleges).

The IPEDS system is, further, not immune from errors in the data that *is* collected. Researchers have, for example, found that zeroes are filled in for missing values in some program-specific graduation rates, artificially deflating a university's performance on student outcomes (Kelchen, 2023). The lack of justification for such handling of missing data (see 5.1.2 above) suggests that this is a processing error rather than a deliberate missing data imputation. Similarly, spending irregularities from flagship state universities have little accompanying information identifying the particular sources for expenditure cuts or increases, reducing confidence in their accuracy (Fuller and Korn, 2023).

None of these errors are severe enough to conclude that the IPEDS data is entirely corrupted or compromised, however these errors raise concerns. A comprehensive and transparent investigation into errors, outliers, missingness, and other abnormalities that affect statistical inferences from IPEDS data (see Section 4.3 as well as Section 5) are highlighted in the specific data elements used from IPEDs, raising further concerns over the quality of the data.

4.2.2. College Scorecard

The College Scorecard, also a product of the U.S. Department of Education, differs from IPEDS in its narrower focus on characterizing the cost and value of higher educational institutions. This spans five areas:

- Cost of attendance, e.g., average net price for students of different household income categories,
- Graduation rate (overlapping with IPEDS), e.g., proportion of full-time degree-seeking students that receive their degree within 6 years of enrollment,
- Employment rate, measured via the linkage of federal loan recipients to de-identified tax records,
- Earnings, measured via the linkage of federal loan recipients to de-identified tax records,
- Average amount borrowed, i.e., federal loan debt, as recorded in the National Student Loan Data System (NSLDS),
- Loan default rate, i.e., of students within various stages of enrollment, as recorded by NSLDS.

With this student-oriented focus, the College Scorecard is a complement to the more institution-oriented variables offered by IPEDS. College Scorecard is used to quantify federal loan burdens by USNWR and Forbes, estimated first generation student graduation rates by the USNWR, net price of each institution by the WSJ, and alumni earnings by Forbes, USNWR, and WSJ.

A key difference with IPEDS is that College Scorecard only captures this information for federal loan recipients. It draws from the same underlying data NCES used by IPEDS combined with other student-level disclosures from the FAFSA loan program, the National Student Loan Data System (NSLDS), and tax records from the IRS.

Although College Scorecard provides high-quality, detailed information for federal loan recipients, one of the main criticisms for using it to draw data elements such as value-added earnings is that it does not

include the non-recipients or recipients of only private loans. This invokes the issue of proxy measurement where an otherwise high-quality measure of one population is used to make potentially misleading characterizations of another population. This point is illustrated when Whitehurst and Chingos (2015) noted that in the 2013-14 academic year, 60% of students enrolled at Stony Brook University held federal loans, while roughly half that percentage of students held loans at nearby private competitor Adelphi University. Yet, they showed, Adelphi's alumni earnings estimated in the Scorecard were about $5,000 lower than Stony Brook's. Rather than a "true" under-performance relative to Stony Brook students, this differential may be explained by an under-representation of its more significant proportion of economically advantaged students in the data. A more thorough review of the specific data elements used in college rankings from College Scorecard is done in Section 4.3.

4.2.3. Surveys

Surveys in college rankings are evaluative tools that gather perceptions about the quality of institutions from academics, industry professionals, and students. These surveys ask respondents to assess the academic and research standards of universities, the employability of their graduates, and other indicators of institutional prestige and performance. The use of reputation surveys, in particular, is based on the premise that those actively engaged in higher education and related industries are well-placed to judge the relative merits of institutions, offering insights that might not be captured through quantitative data alone (Morse & Brooks, 2023; QS Quacquarelli Symonds Limited, 2015).

USNWR, WSJ, QS, and THE each employ surveys to varying extents and for different purposes within their overall evaluation frameworks. USNWR employs reputational surveys extensively, targeting university administrators and faculty, asking them to rate the academic quality of peer institutions. Similarly, WSJ incorporate feedback from student surveys exploring engagement and teaching quality among other characteristics of their environment (Carr, 2023). QS World University Rankings also utilizes academic reputational surveys, directed at academics to assess teaching quality and graduate employability (QS, 2015).

Surveys are also assigned large categorical weights across the four systems. USNWR allocates 20% of a national university or liberal arts college's total score to peer assessments from university administrators, aiming to measure academic quality (Morse & Brooks, 2023). QS amplifies the role of reputation by dividing 40% of an institution's score between academic (30%) and employer (10%) reputation surveys, seeking to identify research excellence and graduate employability (QS Quacquarelli Symonds Limited, 2015). Similarly, THE relies on a global academic reputation survey contributing 33% to rankings, split between teaching and research reputation, intending to capture global scholarly esteem (Ross, 2023). The reliance on surveys underscores the nuanced challenges of balancing subjective perceptions with the objective assessment of institutional quality and impact.

While the use of surveys in college rankings may be popular, surveys themselves are fraught with issues identified in the earlier discussion of total survey error stemming from the inherent subjectivity and potential biases of their questions. The subjective nature of responses, influenced by factors like media representation, historical prestige, and individual experiences, may not reflect an institution's quality or performance, however it may be conceptualized (Sauder & Espeland, 2009). This subjectivity is further complicated by a feedback loop, where prestigious institutions benefit from higher scores that reinforce their status, often to the detriment of smaller or emerging institutions that might excel in areas not captured by general perceptions of quality (Altbach, 2015).

Reputation-based evaluations may broadly impact institutional behaviors, encouraging universities to focus on enhancing their perceived value through marketing and branding efforts rather than improving the actual quality and outcomes of the education they provide. The result is an allocation of resources away from academic programming and student services and towards efforts aimed at boosting survey scores (Bastedo & Bowman, 2010). Other attempts to "game" the system may include leveraging faculty and alumni to participate in surveys or otherwise influencing responses to improve their rankings. This manipulation not only challenges the validity of the rankings but also questions the ethical implications of such strategies on higher education's broader mission (Hazelkorn, 2015).

Additionally, the global nature of reputation surveys, in particular, introduces cultural and regional biases, with English-speaking, Western institutions frequently dominating top rankings positions, thereby disadvantaging institutions in non-Western countries or those not primarily using English for instruction (Marginson, 2014).

Besides the possibilities of measurement error, surveys may suffer from sampling and coverage errors if few or only select students are invited to take the surveys. They may further suffer from nonresponse error if students with stronger positive or negative evaluations are more likely to provide an answer, skewing perceptions in the sample away from the true distribution of the broader student body.

In another regard, ranking providers do take care to mitigate coverage error by updating college contact information annually. A synchronized collection schedule may additionally reduce issues of data lag across institutions. both QS and USNWR receive academic reputation respondents contact information directly from the institution, potentially leading to biases related to the incentives or interests of the institutional representatives (Morse & Brooks 2023; QS World Rankings, 2015). Conversely, THE receives their contact information from a third-party database, Scopus, reducing the potential for institutional influence on response and increasing the credibility of their data (Ross, 2023).

The lack of granularity also poses challenges in survey data by potentially obscuring significant variations in reputation across different programs or disciplines within the same institution. Efforts to address this challenge, such as academic subject-specific rankings within the QS World Rankings and U.S. News & World Report, offer more detailed insights to promote a multidimensional understanding of institutional quality (Morse & Brooks 2023; QS World Rankings, 2015). However, methodologies are still insufficient in reporting granularity to the overall institution rankings, and without transparency about the composition of respondents, these concerns remain.

Despite these challenges, reputation surveys, in particular, remain a staple of global university ranking systems, primarily because reputation, albeit difficult to quantify, matters in academia. Institutions well-regarded by peers and employers can attract top faculty and students, securing a virtuous cycle of excellence and visibility. Some argue that while flawed or subjective, higher ed public opinion surveys offer insights into intangible aspects of educational quality, such as the strength of alumni networks, the value of degrees in the job market, and the global influence of research produced by institutions. Improvements in survey methodologies, increased transparency, and efforts to mitigate bias could enhance the reliability and utility of reputation surveys in global college rankings (Hazelkorn, 2015). Given their acquisition via survey instruments, the resulting measures of perceived reputation or quality must be interpreted with an understanding of the potential for total survey error. At present, it is unclear how USNWR or any other system attempts to mitigate these given the sparsity of methodological details (Morse & Brooks, 2023).

4.2.4. Third-Party Vendors

Two prominent third-party organizations provide supplementary data for (at least) two of our exemplar systems: public policy think tank Third Way provides estimates of earnings premiums, or "value-added," for colleges (used by Forbes and used to inform WSJ's methodology) and technology company PayScale provides estimates of alumni earnings across colleges (used by Forbes).

These data are used to supplement, rather than place-hold, for the aforementioned governmental data sources capturing the same quantities. The Forbes methodology report acknowledges that "neither data source, [College Scorecard or PayScale], is perfect," noting that "PayScale relies on self-reported survey data, which can result in skewed information – graduates that are employed and happy with their earnings can be more likely to respond." Notably, Third Way relies on the publicly available College Scorecard, ensuring that its output measures can be traced back to federally regulated data that adheres to established documentation and methodological standards.

While leveraging multiple data sources can typically strengthen measurement precision and fill in for missing data, it is not straightforward to assume that the imperfections in two datasets are 'cancelled out' when they are combined. In fact, coverage error or other quality issues in one source may even negate the benefits of the other. Ranking providers themselves cite the dangers of incorporating measures from different sources: in 2013, Robert Morse, the Director of Data Research for U.S. News went as far to say that USNWR did not consider the usage of PayScale at all "because it did not consider the data adequately comprehensive or reliable" (Stewart, 2013). If instead – like the data offered by the IPEDS system – the federal government required nationwide reporting of income, Morse stated that USNWR would be more likely to incorporate it.

Further confidence could be instilled in these sources if their precise complementarities (or lack thereof) with their administrative counterpart could be ascertained (e.g., added coverage). While the institution-level data from both of the aforementioned vendors is searchable, only Third Way produces a detailed methodology report and a downloadable file of each institution's estimates (see Itzkowitz (2020) and updates at Rounds (2023)).

4.2.5. Other Public Sources

With the data they offer, third-party vendors may exercise wide discretion in the public release of methodological information, in some cases to protect sensitive or patented technologies. In contrast, data from other public (particularly government) sources usually must comply with certain disclosure requirements and established methodological practices. These requirements apply, for instance, to data from the American Community Survey and the U.K.'s Higher Ed Statistical Agency, which are used by the WSJ and QS to estimate high school graduates' salaries and faculty-student ratios, respectively. The National Center for Science & Engineering Statistics' Survey of Doctoral Recipients is another example, used to estimate future academic achievement in the Forbes ranking system.

Forbes' usage of academic fellowships (e.g., Fulbright, Truman, Goldwater, Rhodes) similarly confers this transparency benefit in that the nomination processes are publicly known. Issues with data quality lie not in the construction or disclosure of these lists, but with their narrow conceptual scope and, thus, potentially problematic usage as a proxy for the general academic achievement of an institution's students.

4.2.6. Direct Disclosure from Institutions

The last major bucket of data sources is the higher-ed institutions themselves. For instance, every year universities are asked to report a whole host of statistics directly to the USNWR: recent Pell graduation rates, summaries of incoming students' standardized test scores, student retention rates, graduation rates, and full-time faculty salaries. The QS system requires universities to disclose information across numerous surveys, including a survey on sustainability practices. In other cases, as with THE, institutions must provide subject-level student outcomes and sign off on their usage in the rankings.

An advantage of direct disclosure from a data quality standpoint is the ability to gather more recent statistics from the recent academic year, thus corresponding to the timescale of other statistics (mitigating the issue of data lag) and more accurately representing each institution's recent accomplishments. There is, however, latitude and discretion given to administrators in the interpretation of requested items. Columbia University, for example, was found to have exaggerated (in its favor) its faculty-to-student ratios, undergraduate class sizes, and faculty educational attainments (Thaddeus, 2022; Mendoza, 2022). In the realm of environmental/social/governance (ESG) practices, stated preferences on climate governance may not always align with the revealed preferences of academic institutions as measured via carbon emissions. Similarly, large systematic discrepancies have been observed between companies' external stances on diversity, equity, and inclusion (DEI) and their hiring practices (Baker et al., 2024).

There is clearly much reason to exercise caution in the reliance of any of these data sources to quantify features of higher educational institutions. In some cases, the choice of aforementioned datasets exacerbate these concerns while in others lend desirable properties to the measurement process.

4.3. Review of Specific Measures

We turn now to the granular measures produced by the sources reviewed in the previous section. Specifically, we have selected measures that have recently been given significant weight (relative to other measures) in the construction of either category or subcategory measures in many of the ranking systems.

4.3.1. Graduation Rates

Graduation rates serve as pivotal indicators across ranking systems, reflecting student outcomes and potential for social mobility. These metrics vary—including 6-year rates of receiving the intended degree for first-year BA students, Pell Grant recipients, and first-generation students—and are weighted differently across systems.

In the 2024 U.S. News & World Report rankings, the 6-year rate for first-year BA students holds a 16% weight, while Pell Grant and first-generation rates hold 3% and 2.5%, respectively. Furthermore, U.S. News incorporates these rates in modeled predictions of graduation rates, integrating socioeconomic and resource data to refine its assessments. Forbes similarly emphasizes six-year graduation rates, allocating a 15% weight to this metric, delineated between all students and Pell Grant recipients. The Wall Street Journal also employs this measure, underscoring its widespread relevance across major ranking systems. Typically, these data are sourced from the Integrated Postsecondary Education Data System (IPEDS), ensuring a standardized national database.

The application of 6-year rates specifically for full-time, first-year students targets standardized comparisons by excluding transfer and part-time students. While this ensures consistency in tracking a specific student demographic across institutions, sidelining non-traditional student paths notably limits the breadth of the data's applicability. In construct validity terms, this may affect the content validity of the rankings by not representing the experiences of these students – what may also be understood as an issue of proxy measurement.

Moreover, the averaging of data from multiple cohorts (e.g., the 2013 to 2016 entering classes) attempts to smooth year-over-year fluctuations but introduces potential biases through the variable timing of data maturity—earlier cohorts have data that has undergone more cycles of review and potential revision. This is also an instance of data lag: a college may be performing well in a given year on other contemporaneous quantitative indicators considered in its ranking but either be artificially penalized ('dragged down') or boosted ('dragged up') by its 3-year average.

Given that IPEDS reports 4, 6, and 8-year rates simultaneously, there exists an inherent delay of up to two years before such data become available. This delay, alongside the varied data release schedules (e.g., provisional data released nine months post-collection and finalized data following the IPEDS Prior Year Revision system), potentially affects the freshness and relevance of the data used in rankings. Moreover, the granularity of graduation data—including rates for specific subgroups such as first-generation students or race/ethnicity subgroups—offers a deeper, more nuanced view of institutional performance. Yet, these detailed cuts are underutilized in current mainstream ranking frameworks, suggesting an area ripe for future expansion to increase the accuracy of rankings when it comes to representing outcomes for an *entire* student body.

Finally, graduation rates are often used to solve the problem of missing data in *other* measures, but in a ad-hoc manner. Interestingly, U.S. News adjusts its weightings based on the availability of standardized test scores, demonstrating an adaptive but somewhat opaque methodology in response to incomplete data. This flexibility in handling missing data points, such as increasing the graduation rate weighting from 16% to 21% when fewer than 50% of ACT/SAT scores are reported, reveals a pragmatic approach but also introduces variability that may challenge the consistency of the rankings.

Despite ongoing improvements to the IPEDS methodology and data collection practices, transparency around these enhancements remains limited, obscuring the full scope of data reliability and validation processes. Such gaps underscore the critical need for clear documentation and open communication about methodological adjustments and data quality assessments to bolster stakeholder confidence in the derived rankings.

While graduation rates are integral to assessing educational outcomes within college rankings, the methodologies employed, and the depth of data integration reflect varying levels of sophistication and transparency across different ranking systems. This variability underscores the importance of rigorous methodological scrutiny and the need to refine these approaches to enhance both the accuracy and the interpretative value of college rankings.

4.3.2. Student Debt

The incorporation of student debt data within college ranking methodologies addresses a significant concern of higher education stakeholders: the financial burden shouldered by students. Amid escalating college costs and the consequent reliance on loans, student debt metrics serve as crucial indicators of financial accessibility and the potential return on investment (ROI) associated with a college degree.

For example, U.S. News & World Report now allocates a 5% weighting to median student debt upon graduation, an increase from 3%, acknowledging its influence on prospective students' decision-making processes (Morse & Brooks, 2023). Similarly, Forbes factors post-graduation student debt into their rankings, assigning it a 15% weight to gauge the economic value of the degrees offered by institutions (Whitford, 2023).

Both U.S. News & World Report and Forbes utilize the College Scorecard to ascertain student indebtedness. Integrating College Scorecard into ranking methodologies offers standardized, federally collected data, enabling consistent comparisons across institutions. This alignment can sharpen the analysis of post-graduation outcomes, addressing central concerns about debt levels and employment prospects. Yet this choice raises significant issues concerning the relevance and comprehensiveness of the data.

Student debt, as it is reported in the College Scorecard, is among the most concerning usages of proxy measurement. College Scorecard predominantly covers federal student loans and omits private loans, which represent around 10% of all student debt. Given that private loans often entail higher interest rates and less favorable repayment conditions, their exclusion can markedly underestimate the total debt burden of graduates (NCES, 2023), highlighting the dangers of “placeholder” data between populations.

Furthermore, the Scorecard restricts its reporting to graduates only, overlooking individuals who drop out of college. These non-completers frequently face severe financial strains without the economic uplift that a degree typically ensures, thereby skewing debt statistics downwards and potentially portraying an overly sanguine scenario of financial health post-college. The Scorecard also excludes students without aid. This group may rely on different financing methods, such as private loans, personal funds, or institutional aid, none of which are captured in the dataset. As a result, Scorecard reports of graduation rates fail to encapsulate the full spectrum of student financial experiences.

While detailed debt data, including median debt levels and repayment figures, are often available from sources like Scorecard, rankings typically use aggregate averages that might obscure significant intra-group variations. Institutions’ rankings may be influenced disproportionately by averages that do not account for the wide disparities within the student body, a methodological shortcoming that could mislead stakeholders about the true financial implications of attending certain colleges. The U.S. News & World Report relies on the median figure from College Scorecard, thus likely misstating the affordability of institutions with high costs (Morse & Brooks, 2023). Forbes similarly uses median federal loan debt but does not specify if their data includes non-graduates or considers other missing data elements (Whitford, 2023).

Finally, student debt measures from Scorecard and elsewhere does not cover institutions that opt out of reporting student debt data or that offer extensive scholarship programs. This added layer of missing-ness only further skew perceptions of the financial burden associated with college education.

The comprehensive nature and simplicity of IPEDS and the College Scorecard make them invaluable data sources, yet they present formidable challenges for non-specialists given the immense volume and complexity of the information they contain. Further complicating the picture is the retrospective nature of student debt data, which may not accurately reflect current borrowing trends or changes in financial aid policies.

To deliver insightful analyses, the integration of student debt data within ranking frameworks must be coherent and nuanced, balancing debt metrics against educational quality, graduation rates, and alumni

earnings. This integration should be transparent, with clear methodologies for addressing data discrepancies and biases. Without such clarity, the scientific integrity of these rankings could be called into question, undermining their validity as tools for prospective students navigating their educational options.

4.3.3. Value-Added Earnings

In recent years, the use of value-added earnings, or return on investment (ROI), has gained significant traction in major college rankings. Notably major ranking systems such as Forbes and the Wall Street Journal leverage granular data sources such as institutional surveys, proprietary databases, and government datasets to analyze graduates' earnings across various dimensions. The Wall Street Journal compares median earnings among different schools and geographic regions, while Forbes also compares geographic region, as well as a factor on value added based on a graduate's economic backgrounds (Carr, 2023; Whitford, 2023). These rankings aim to provide prospective students and stakeholders with valuable insights into the value added by colleges and universities in terms of their graduates' earnings. However, methodological deficits such as ambiguity in conceptualization and representation of the salary data used pose threats to the construct validity of these rankings.

It is worth emphasizing that there is no authoritative comprehensive source for student earnings. Pending any legislative changes, the federal government is unable to link college students' incomes to their degree status and publicly disclose the resulting data. Individual institutions have no such mandate to do so either. Instead, ranking providers must rely on other federal and private data products, each with their own deficits in quality. Forbes relies on data from the College Scorecard to evaluate the value that colleges and universities add to their students' earnings *for loan recipients* (Whitford, 2023). On the other hand, the Wall Street Journal leverages data from the National Center for Education Statistics (NCES) alongside proprietary surveys from research agencies like College Pulse and Statista to assess graduates' economic achievements (Carr, 2023), though notably these are *sample surveys* rather than estimates for the entire population. Additionally, third-party organizations like Third Way and Payscale provide supplementary data on earnings premiums and alumni earnings, respectively, which are utilized by Forbes and the Wall Street Journal.

The importance of leveraging value-added earnings data in college rankings lies in its ability to offer a tangible measure of the economic benefits that students derive from their education. Incorporating earnings data into the ranking methodology provides a practical indicator of the ROI for students and can significantly influence their decision-making process when choosing a higher education institution (Carr, 2023; Whitford, 2023). Forbes' use of College Scorecard serves as a seemingly appropriate source of ROI as a ranking indicator, but challenges related to representation and use of College Scorecard data can introduce errors that affect the overall validity of the rankings.

Concerns associated with the use of College Scorecard's design are identified in a detailed Brookings report at the time of the Scorecard's release (Whitehurst and Chingos, 2015). Chiefly, the salaries of federal loan recipients may not – as iterated earlier – be representative of non-recipients (or recipients of the Pell Grant or only private loans) whose salaries are excluded from the data. Also, College Scorecard does not have a granular approach to program specific salary data, only the institutional level.

In comparison, the Wall Street Journal leverages data from proprietary self-administered surveys from research agencies like College Pulse and Statista to assess graduates' economic success (Carr, 2023). Incorporating external data sources introduces complexities related to criterion validity, as the alignment between the chosen measures and the underlying construct of economic success must be thoroughly

evaluated to maintain the credibility of the rankings. And while these sources detail a more thorough vetting process on data quality, the lack of transparency and accessibility of the data itself prevents a thorough review of the data and the rankings output compared to their stated purpose, raising concerns about the overall methodological integrity and validity of the findings.

The credibility of rankings based on earnings data is contingent upon the accuracy and reliability of the information presented. Forbes and the Wall Street Journal employ validation processes and quality assurance measures to maintain the integrity of the data used in their rankings, though challenges persist in ensuring the accuracy of self-reported data sources like Payscale. As with student debt figures. reporting aggregate summary statistics, such as the median, ignores the considerable variation in salaries between degree programs. Mean or median salary reports also may not be informative or even distorting when presented without the percentage of students that have no loans. As in the case with other conceptual dimensions being measured by rankings, it is crucial that value-added earnings are supplemented with other measures of 'value' to achieve content validity.

By addressing methodological challenges and enhancing the transparency and reliability of data sources, ranking systems like Forbes and the Wall Street Journal can strengthen the validity of their value-added earnings metrics, fostering more robust and meaningful insights for stakeholders. Alternatively, legislative action can be pursued to federally mandate the linkage and disclosure of *all* students' degree and earnings outcomes to provide the highest quality data to the consumer public and the ranking systems alike.

4.3.4. Fellowships

Publicly available records of prestigious fellowships and scholarships are most notably included in the Forbes ranking system as a measure of academic success. While the data quality of fellowship recipients is considered reliable due to its public availability, incorporating these metrics into ranking systems can introduce challenges related to conceptual scope and validity.

For example, in the Forbes 2024 Top Colleges list, academic success constitutes 10% of the total. The academic success comprises an aggregation of two subcategory measures: the number of recent graduates who have won prestigious academic fellowships, such as Fulbright, Truman, Goldwater, and Rhodes scholarships (5%), and the number of recent alumni who have gone on to earn doctorate degrees using data from the National Center for Science and Engineering Statistics (5%) to determine the average number of alums who earned a Ph.D. over the last three years. While it is helpful that multiple, theoretically distinct variables are used to measure academic success, this measure fails to capture the *entire* distribution of academic performance at the university. This bears repeating a central theme in our report: multiple measures must be brought to bear to accurately and comprehensively cover broad conceptual dimensions such as 'performance'.

One of the key validity concerns in utilizing fellowship data is the relevance of these awards in measuring academic quality. The specific focus and criteria of prestigious fellowships like Fulbright, Truman, Goldwater, and Rhodes scholarships may not align with the comprehensive definition of academic success across diverse fields and institutional goals. Different fellowships cater to distinct purposes and disciplines, leading to a narrow representation of academic quality within the rankings that only recognize the most decorated students.

Moreover, the coherence of using multiple fellowships with varying objectives raises questions about the alignment of the data element with the broader ranking framework. Each fellowship follows unique

selection criteria and standards, contributing to different levels of competitiveness and rigor in recipient selection. The lack of clarity in the Forbes methodology regarding the rationale for selecting specific fellowships to assess academic performance highlights potential coherence issues in the evaluation process.

The temporal aspect of fellowship awards presents another challenge to the validity of rankings. Fellowships like Fulbright and Goldwater may honor recent college graduates or undergraduate students, respectively, at different stages of their academic career. As a result, incorporating fellowship data from different academic stages into a single metric may not effectively align with the academic successes of the universities being evaluated in a given ranking year – another instance of data lag.

While the credibility of fellowship data is generally high due to the reputation and selectivity of these awards, the methodological approach to integrating multiple fellowships into rankings should ensure coherence with other measures and relevance to the intended conceptual dimensions. Ensuring that fellowship data aligns with the broader goals of the rankings and accurately reflects the diverse dimensions of academic success is pivotal for maintaining the validity and robustness of the ranking systems.

4.3.5. Financial Resources

The application of financial data in college rankings methodologies is intended to deliver a multifaceted view of institutional fiscal health, resource allocation, and investment in educational quality. However, the integration of such data brings to the forefront a variety of validity challenges that may compromise the integrity and utility of the rankings produced.

Financial data is pivotal for assessing institutions' stability and their capacity to support academic programs and student services. USNWR, for example, considers financial resources as a crucial criterion, focusing on the average spending per student on instruction, research, and student services (Morse & Brooks 2023). However, the current methodologies employed for collecting financial data, such as those by IPEDS, underscore significant disparities in reporting, particularly influenced by institutional structures. For instance, larger institutions with multiple sub-entities, or "children institutions," report financial data separately from their "parent institutions." This separation often leads to underreported financial resources for the primary institution, skewing comparisons with institutions that report consolidated financial data.

Although financial information is derived from federal databases or institutional reports, the accessibility and detail presented publicly are inconsistent. USNWR and other ranking entities frequently outline their methodologies yet leave specific data points and calculations opaque, muddling stakeholders' understanding of how exactly financial metrics influence rankings (Morse & Brooks 2023). Moreover, the inherent delay in financial reporting — given that data often reflect the previous fiscal year — adds another layer of complexity, potentially misrepresenting an institution's current financial health.

The granularity of financial data is another pressing issue. While broad metrics provide a high-level view of institutional spending, they often fail to capture the nuanced allocation across various programs or services that directly impact the educational experience. The lack of disaggregated data hampers a comprehensive understanding of how resources are deployed within institutions, which is essential for prospective students and faculty considering different programs.

The accuracy and reliability of financial data hinge on stringent data collection and verification processes, along with adherence to consistent financial reporting standards. Variations in accounting practices and fiscal years, among other factors, complicate direct comparisons across institutions. Standardization bodies like the National Association of College and University Business Officers (NACUBO) recommend guidelines to align financial reporting practices (NACUBO, 2021). Nonetheless, adapting these guidelines across diverse educational institutions remains complex, often necessitating further normalization adjustments for fair comparisons.

The scientific integrity and credibility of financial resource data, as managed by entities like IPEDS, are generally robust, given the federated oversight and multiple validation steps involved in the data collection process (NCES, 2023). Even so, the broader application of these data within rankings methodologies must be scrutinized to ensure that financial metrics do not overshadow or skew the qualitative assessments of institutional performance.

To recap, while financial resource data play a fundamental role in evaluating colleges and universities within ranking systems, numerous challenges related to the validity, accessibility, timeliness, and granularity of these data must be addressed. Enhancing transparency and refining granularity are imperative steps towards more credible and balanced college rankings. Until these validity issues are adequately addressed, stakeholders should remain cautious in interpreting ranking outcomes based largely on financial metrics.

5. Review: Methodologies

In this section, we turn to the final stage of the measurement process across all ranking systems: transforming the data mentioned in the previous section through various statistical procedures into numeric quantities that are then cohered into a ranking for each college in the system's universe. This is the crucial process of operationalization. We focus on five major buckets of issues across the methodological choices made in this process, starting with the most critical issue: the choice of an ordinal ranking itself as the measure.

5.1. The Equal Intervals Problem

To the unsavvy consumer, ordinal (or ordinal-like) scales may *appear* to imply equal value intervals between units. For instance, one might assume that the gap in quality between colleges ranked #2 and #3 implies the same difference in quality between colleges ranked #10 and #11. In general, this is not true of ordinal variables since they do not imply anything about the underlying criterion used to rank the units (Stevens, 1946).

On a cursory inspection, we find this to be true in our college ranking systems: in the WSJ 2024 ranking, the gap in the overall score (on a scale from 0-100 and used to directly determine rank) between the #2 and #3 universities (Yale and Stanford) is 0.2 while it is 2.0 between the #10 and #11 universities (Babson and Swarthmore). Thus, according to the very scale used to generate the WSJ rankings, there are varying degrees of differentiation between colleges merely a rank apart.

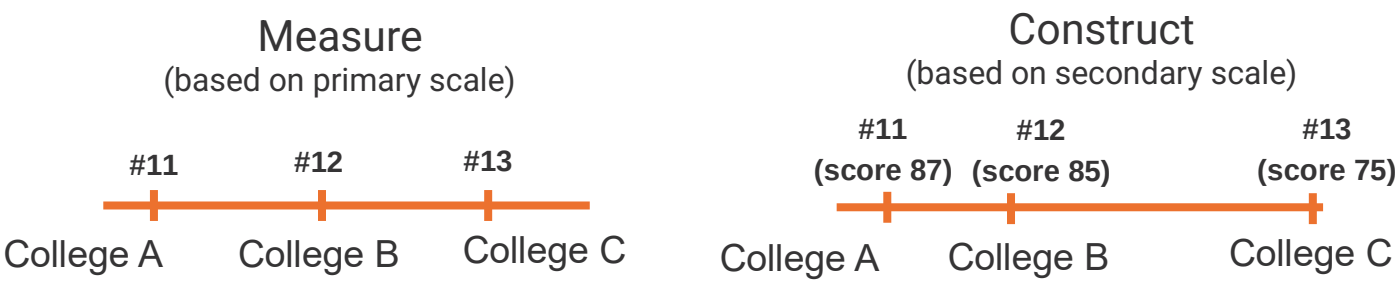


Figure 3. Illustration of the Equal Intervals Problem in a Fictional Ranking

The above figure illustrates this dilemma more clearly with a fictional (hyperbolized) example. The appearance to consumers is the primary scale (i.e., the ranking itself) which is shown on the left: three institutions ranked one after the other, implying relatively equal gaps in the underlying construct captured by rankings. The true values of construct may be approximated or even exactly captured by the secondary scale (i.e., what is directly used to construct the ranks) on the right: colleges #11 and #12 are much closer apart in the construct than #12 and #13.

Thus, the usage of rankings itself is subject to a fundamental issue of construct validity – namely, differences in where institutions lie along the underlying construct are over-simplified. The result runs deeper than a technical concern of measurement: consumers may be misled to believe that colleges are more (and less) differentiated than they actually are.

5.2. Subjective Weights

In all cases, the underlying quantitative score that is used to construct the ranking is itself a weighted combination of different component measures. Across our exemplars, these weights are based on subjective evaluations of conceptual importance rather than values derived empirically or through some defensibly objective criterion -- except for the QS World Universities ranking system where weights are determined via surveyed student priorities. Even in the case of QS, it is unclear how student priorities were aggregated (e.g., average, median, majority vote) and translated into weights. The lack of methodological documentation, as in the case of data documentation, poses a substantial barrier to evaluating the robustness of the selected weights.

Some systems justified *changes* to weights, but not the precise number while others did neither. A recent trend (USNWR 2024, WSJ 2024, and Forbes 2024) has been to increase weights from previous rounds allocated to student 'outcome' categories (e.g., earnings) relative to educational experience or other 'input' variables (e.g., standardized test scores). In other cases, weights assigned to specific category measures in previous years were redistributed to accommodate the introduction of new categories.

A 'weight-and-sum' approach to combining measures to represent an unobservable construct is inappropriate since, under the latent variable interpretation, the weights between observable indicators are themselves conceptually unobservable (Guarino, 2005). Furthermore, the usage of deterministic, manually constructed weights ignores fundamental uncertainty in the measurement process itself (Guarino, 2005). If one combines measures of different *dimensions* (i.e., the concepts underlying the category scores discussed in Section 2) of a construct, the weights could be determined by some statistical procedure relating the dimensions to some independent measure related to the concept – for example, using regression analysis to determine how well each *dimension* (i.e., category score) predicts a

measure of reputation. Using this procedure would also have the favorable side effect of determining the criterion validity of the overall measure.

Alternatively, the weights may be determined on the basis of how 'informative' or 'discriminative' the component measures are: measures with little variation across universities are less helpful in capturing variation along the construct of interest, are less helpful in differentiating one university from another, and, therefore, should be given a lower weight. Techniques such as factor analysis can help identify such weights.

5.3. Categorization and Transformation

Beyond the usage of subjectively determined weights, college ranking providers make other subjective choices in the transformation of raw datasets into operational quantities. While subjectivity does not necessarily undermine the integrity of measurement, it creates an added layer of discretion; that is the provider can make a choice that favors some preferred, pre-determined outcome. In many cases, to the credit of our exemplar systems, we find that these subjective choices are justified by some criteria independent of specific institutions.

Classification

Certain measures used in the construction of rankings required classification decisions, that is categorizing features of institutions (e.g., degrees) into discrete categories (e.g., discipline). In some cases, this classification appears to be done by subjective judgment – or, at the least, a procedure that is insufficiently explained. For example, QS 2024 and THE 2024 both use custom classifications of "narrow subjects" which are used to determine the offerings from each university in the inclusion/exclusion stage. In other cases, we recognize the usage of external 'crosswalk' datasets to perform the classification, although even here there is discretion in the choice of crosswalk.

As an example, USNWR 2024 calculates the portion of degrees that are STEM out of the total degrees awarded in the graduation performance model using the US Department of Homeland Security's (DHS) STEM Designated Degree Program list – apparently, because it is inclusive of "specific STEM degree tracks in nontraditional STEM fields, such as business statistics and digital communication and media." The DHS STEM program list, however, originated as a criterion for extending the visa status of foreign students. It is unclear whether USNWR's intention in using this list rather than others (e.g., the Classification of Instructional Programs from the Department of Education) is to reap some methodological benefits from a more holistic STEM classification or to reward institutions that have a more inclusive scope.

Normalization

Many indicators used across the systems are on different scales; if nothing is done, this will – among other things – implicitly upweight certain indicators with larger mean values and create undue influence for measures with relatively high variance. To prevent this, category scores can be normalized (or standardized, used equivalently here) to equivalent ranges.

The benefit of this is to make otherwise disparate data with different means, variances, ranges, and substantive 'units' comparable on a common scale. The drawbacks include the loss of interpretability of the original data, misleading transformation of outliers (e.g., misleadingly closing large gaps between institutions), and incorrect/unverified assumptions about the normality of the underlying data (e.g., if z-scoring and a cumulative distribution function is used without demonstrating evidence of normality). The

reported normalization and standardization decisions for each of the five exemplar systems are noted in Appendix Table A6.

Truncation

Nearly all of the five exemplar systems only assigned numeric rankings to a select top percentile of universities (within the original sample) and assigned lower-ranked universities to tiers or omitting them from ranking altogether. Thus, the ordinal scale is nearly always shown in partial form: units that fall below a certain threshold are either truncated (removed) from the scale entirely or collapsed into intervals. Because of these transformations, college rankings are not truly ordinal, but better described as 'semi-ordinal' or 'partially ordered categorical variables' (DeVellis and Thorpe, 2021). Reported truncation decisions by our five providers are described in more detail in Appendix Table A7.

In one way, the operational decision to truncate rankings bolsters the methodology by acknowledging the uncertainty in precisely ordering institutions at the lower tail end of the ordinal scale. Institutions on the lower end of category scores across our systems may disclose fewer variables to the providers, receive fewer audits on the information they do disclose and, hence, may be difficult to differentiate from each other. The trade-off is a reduction in overall information provided by the measure at the tail end, though there is less conceptual utility to differentiate between colleges at this tail in the first place. For college-goers interested in institutions at this end of the ranking scale, specialized observable characteristics of these colleges rather than a broader concept of institutional quality encompassing factors like prestige are likely to be more useful.

Adjustments

Adjustments are often performed to a variety of measures to compensate for errors of representation. For example, statistical adjustments to survey-based estimates are usually made after a survey is conducted to reduce coverage, sampling, and non-response errors using known information about the target population, frame population or response rates of the sample (Groves, 2005). Still, errors in an adjusted statistic – aptly called *adjustment error* in the Total Survey Error framework -- could result from differences between the adjusted statistic and the true but unknown population value represented by that statistic due to inaccuracy in the known information about the target population. College Pulse, the most prominent survey across our providers, applies "a post-stratification adjustment based on demographic distributions from multiple data sources, including the 2017 Current Population Survey (CPS), the 2016 National Postsecondary Student Aid Study (NPSAS), and the 2021-22 Integrated Postsecondary Education Data System (IPEDS) to rebalance the sample based on a number of important benchmark attributes, such as race, gender, class year, voter registration status, and financial aid status." It also uses an iterative proportional fitting (IFP) process to balance the distributions of all variables. College Pulse states that it trims weights to prevent individual interviews from having too much influence on the results, although it is unclear what kind of trimming strategy is used and the effects of this trimming on the quality of the adjusted statistic.

Moreover, error adjustments at the *very* tail end of the measurement process – to already published rankings – are not uncommon. For example, the 2024 USNWR ranking had to correct the scores of 213 schools after publication due to "a code anomaly" in the code collecting input data. The 2024 QS ranking had two institutions (Abu Dhabi and Birkbeck) notify them after publication that they had provided incorrect numbers or been excluded respectively; QS responded and amended the ranking. A number of other similar errors were corrected; 5 institutions corrected before publication and 3 after publication (QS notably transparently posts about their correction process). Although these adjustments are important to

increase transparency and correct errors in the operationalization of a measure defined in a particular year, there is a risk that such errors may unintentionally undermine consumer trust in the ranking system altogether.

5.4. Uncertainty

There are many sources of uncertainty in the measurement process (see earlier discussions of the Total Survey Error framework). One source that affects statistical estimates based on surveys or other data sources is measurement error. We recap this briefly now.

There are two types of measurement error to consider in any measure. First, there is systematic measurement or systematic deviation in responses from the true value due to poor measurement, such as misreading or misunderstanding of a survey question or the exclusion of low-income students when using data about Pell grantees. This tends to manifest as bias, or a uniform deviation from the true value. Second, random measurement error can result from variability across measurement occasions (e.g., rounding of dollar figures for monetary variables or yielding different answers to the same question at the survey respondent level).

Measurement error, moreover, can stem from many kinds of measurement procedures used in the construction of the rankings: models of unobserved quantities (e.g., economic value-added), surveys of public opinion (e.g., student sentiment), methods for accounting for missing data (e.g., graduation rate substitution for standardized test performance or positing a statistical model from which imputed values are generated), or the creation of weights to combine different measures.

In each of these cases, the standard statistical practice is to construct *confidence intervals* that quantify the degree of uncertainty around a particular measure given:

1. Conceptualization of the 'true' value either (a) observable in the population from which the sample is drawn or (b) fundamentally unobservable with real-world implications, and

2. Assumptions about the mathematical form of both systematic and random measurement error which may involve related assumptions about missing values, sampling processes, response mechanisms, etc.

As we have described in a previous section, truncation of rankings is one way to signal uncertainty – rather than assign an institution to the exact ranking produced by the operationalization, one brackets it in a tier of possible rankings. This, however, is an ultimately ad-hoc, subjective, and imprecise method for conveying uncertainty. More dire, it may *under-estimate* the amount of uncertainty resulting from the measurement procedure altogether.

Uncertainty associated with statistical estimation of individual measures affects the ultimate rankings. The figure below illustrates how the absence of confidence intervals, as in the case of the usage of ordinal ranks itself, can cause misperceptions of differentiation between colleges. On the other hand, the visual display of uncertainty intervals allows for the appropriate inference that it is not possible to exactly place the three institutions in question at the exact ranks shown on the left.

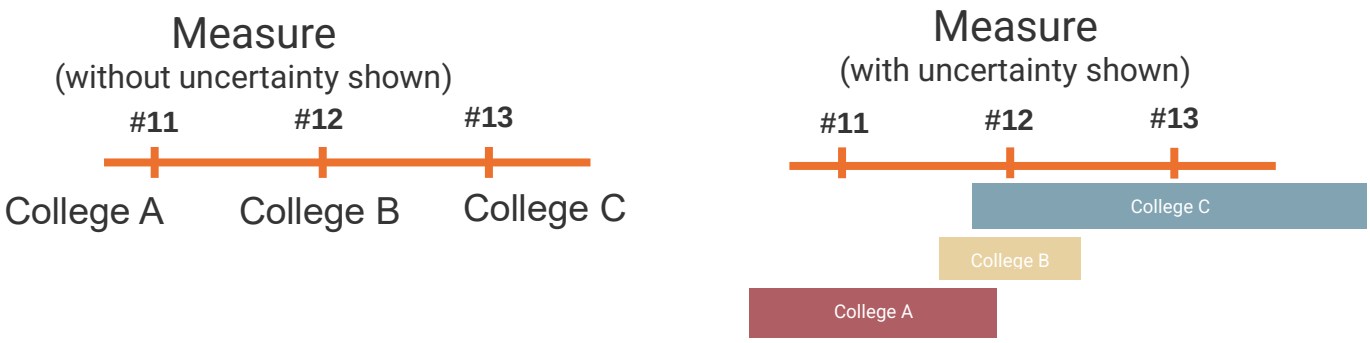


Figure 4. Comparison of the Current Presentation of Rankings (Left) with Possible Presentation Including Uncertainty Intervals (Right)

Once uncertainty is quantified, analysts may take steps to reduce it. Surveys, for instance, might be provided to a random sample of students from each university who are drawn to represent the entire university student body, which is known to reduce the uncertainty associated with sample estimates of the target population. Even if a survey could be delivered to every student, it would generally still be inappropriate to ignore the component of *fundamental* uncertainty in measures of preferences and other variables capturing unobservable concepts. Here, better justified measurement models based on a keen understanding of the data-generating process can help to accurately quantify said uncertainty; usage of explanatory factors can, similarly, reduce estimates of uncertainty Similar logic holds for data sources such as administrative data that might reflect characteristics of the entire population. See Elliott et al. (2006) for a full discussion of these points in the context of profiling health care institutions.

To summarize, each input to a college ranking has uncertainty associated with it, and while that uncertainty propagates to the reported college ranking, albeit in a complex way, it is not included in rankings reports. To date, we have not seen the usage of confidence intervals or other uncertainty statements in any instance across the five exemplar systems. However, uncertainty estimation is an area of significant focus in value-added modeling of teacher performance and healthcare provider profiling, and it has received attention in the context of ranking educational institutions by Goldstein & Spiegelhalter (1996) who show rankings may over-differentiate between institutions that are otherwise indistinguishable when accounting for the uncertainty in ranks.

6. Conclusions and Recommendations

For years, significant attention has been paid to the limitations of rankings from many stakeholders, including those in academia, nongovernmental organizations, and popular media. Despite the attention resulting in proposals to address these limitations, the most influential rankings raise methodological concerns.

In this report, we employ statistical and measurement/psychometric concepts to address the key question of what exactly is measured by the rankings. We frame our discussion in terms of construct validity, and thus review conceptual, data, and methodological factors that pose challenges to the rigor of college rankings. These factors closely align with three of the four dimensions of the Berlin Principles on

Ranking Higher Education Institutions, which were developed in 2006 by the International Ranking Expert Group to provide a framework for the development and use of rankings of higher education institutions in hopes of stimulating future improvements (see Appendix A3). Our report also touches upon and has implications for the fourth dimension of the Principles: The presentation of ranking results. The information we provide should be useful to those adhering to these Principles or to those who are more generally interested in understanding the strengths and limitations of college rankings.

Nearly every aspect of college rankings suffers from a lack of construct validity. This starts with the conceptualization of what is measured, and then extends to *how* concepts are measured in data elements drawn from various data sources, and finally to how the data/information are processed and presented using various methodologies. Improving college ranking systems will require addressing all three areas.

Based on the evaluations in this report, we now present seven possible (not mutually exclusive) pathways for the methodological advancement of college ranking systems.

6.1. Empirically validate existing rankings

A direction researchers may consider is to conduct a psychometric analysis to evaluate construct validity properties introduced in Section 3 such as scale reliability and concurrent validity of the college ranking systems. Researchers could replicate one or more of the exemplar systems (either partially or in full, across consecutive years or a longer time period) to assess their scale reliability and/or evaluate the correlation in rankings within a given year, which could be extended to measuring correlation within rankings across years to demonstrate predictive validity. Correlations of rankings could also be examined for key outcomes. While we would expect strong correlations of rankings over time, it is less clear whether findings would agree with those from studies of professional programs or from earlier time periods given recent changes in the employment market and stakeholder priorities.

The application of these measurement concepts to the development of college rankings appears to be limited, at least in the five influential exemplars examined in this report. Some attention has been in the literature given to psychometric analyses to examine the reliability and validity of certain ranking systems, as noted in Section 3.4. The stability of institutional characteristics is a requisite assumption for some of these analyses which is generally reasonable, but becomes more challenging to assume when the input criteria for ranks change over time and recent changes to major rankings to give more weight to socioeconomic concerns of students. For that reason, substantial attention must still be devoted further upstream in the measurement process: to the question of what the ranks actually reflect and whether the chosen (or other) inputs are adequate to capture it. As we discuss in 3.7, identifying the type of construct (reflective versus formative) has further implications for the types of validity tests (and inherent validity issues) to be considered.

There is a need to build (either from the ground up or from existing efforts) a clear and unambiguous conceptual framework for the scientific construct underlying rankings, its various conceptual dimensions, and the measures that map onto those dimensions. At present, the college ranking systems do not explicitly do this, requiring the consumer to guess what the framework may be from unclear descriptions and measurement choices made without explanation. Thus, future work could focus on developing a more robust conceptual framework for composite construction with clarification on its type as well as greater attention to the elements of construct validity and measurement illustrated in Figures 1-2.

Alternatively, rather than trying to combine disparate measures into a single index at all, ratings for dimensions of 'quality' can be reported entirely separately with the underlying summary statistics so that institutions' actual data can be compared. That is, rather than present an aggregate measure of a single broad construct, college ranking providers may elect to only present the disaggregate measures of multiple specific concepts.

6.2. Establish robust quality and transparency standards

Even if a clear conceptualization and empirical validations existed, the value of the resulting measures would still be compromised by using data not suited to accurately reflect the concepts. Hence, we advocate that college ranking systems adopt a formal quality framework to guide data quality assessments as well as the disclosure of these assessments to bolster confidence in stakeholders. Survey data quality considerations have long benefited from employing the Total Survey Error framework (see Appendix 2) to clarify issues of measurement and representation of a target population.

In the development of such quality standards, we would highly recommend incorporating elements from (or adopting in bulk) the data quality framework from the Federal Committee on Statistical Methodology (FCSM). The FCSM Data Quality Framework was established by FCSM's Data Quality Analysis Working Group to provide practical information on identifying and reporting data quality for federal agencies and other organizations. Data quality is "the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust." The FCSM framework focuses on evaluating data quality in the context of statistical use and decision making. The FCSM Data Quality Framework (Figure A2) has three broad domains: utility, objectivity, and integrity, which are specific areas where data quality can be considered. See Appendix 3 for more details on these domains and a graphical overview of the framework.

Quality standards drawn from the TSE or FCSM frameworks could be accompanied by a set of reporting standards on measurement considerations we have identified in this review (e.g., imputation and prediction quality, accessibility of all underlying data sources, and standard errors associated with subcategory measures).

Beyond the procedural realm (dubbed 'operationalization' in this report), college ranking systems could improve their adherence to transparency at all stages of their methodology. We advise that college ranking systems work with universities and third-party vendors to make the detailed methodology and data that feed into college ranking systems publicly available. This is especially true when public opinion surveys (e.g., reputation surveys, student evaluation surveys) are used. We have proposed using the total survey framework to evaluate potential introduction of error at every stage in survey development and data collection. Implementing such a common framework to ensure high data quality across all universities that submit data to the ranking system will be extremely useful. The same level of transparency is also recommended for methodological decisions that go into developing the ranking system.

6.3. Consider ratings as an alternative to rankings

An inherent methodological challenge in developing rigorous college rankings that leads to its misuse and misinterpretations is that college ranking systems combine measures of several dimensions into one index, which is then treated as an ordinal scale. As we have shown, the distance between each level of the scale is assumed to be constant and is used for ordered rankings (dubbed the 'equal intervals

problem'). We show numerous methodological issues as to why this practice is unfit. While ranks are ordered, the distance between the ranks is not constant. Users of the rankings cannot accurately measure the distance between institutions that may have ranks close together or quite far apart. Indeed, many providers such as The Wall Street Journal (WSJ) include the constituent index (or the secondary scale by the definitions in this report) used to construct the ranking. However, the ranking itself is clearly the focal point of interest for both the providers and their consumers. We also note concerns that are well-known in the educational statistics literature about variability of rankings.

Hence, we may consider abandoning the idea that we can meaningfully rank many complex institutions. Instead, we might elect to rate institutions on a limited (ordinal) scale of 5 to 7 points with qualitative terms such as "poor" to "excellent". While those in the higher categories will be judged "better" than those in the lower categories, there will be no pretense that we can make the distinctions implied by ordered integer rankings. Ideally, the categories of such a rating system would reflect meaningful differences, such as institutions outside of the top category being different in a statistical sense from those in the top category.

6.4. Improve the visual presentation of the final measure

A risk with a transition to limited scale ratings may be best described as the 'cliff's edge' problem: a small change in individual measures may bump an institution down into a lower tier. This dilemma would exacerbate the effects of otherwise small errors in measurement (e.g., rounding of dollar figures, low sample size in surveys). Rather than *reduce* 'gaming' incentives, a rating system might *magnify* the effects of gaming: some colleges may boost their tier with small adjustments to their disclosed statistics.

One solution to mitigate the cliff's edge dilemma is to provide more, rather than less, information about the resulting scale under a rating system. For example, providers may be obligated to visualize colleges on the underlying *continuous scale* used to construct ratings to demonstrate how close or far apart institutions are in the underlying quality of the scale. Providers may also be obligated to properly quantify the uncertainty in each assigned rating and visualize it accordingly using confidence intervals around each institution on this spectrum. These two visual elements could still be paired with 'bracketing' elements that distinguish institutions that significantly (in a statistical sense) pass the cut-off for each tier.

Altogether, we believe a more enriched visual scale along these lines with *ratings*, rather than *rankings*, as the top-line metric would provide more meaningful, less misleading differentiation between institutions.

6.5. Consider personalizing college rankings

The lack of objective criteria for assigning weights to different categories means there is a great degree of arbitrariness in the overall scores and, thus, the rankings themselves.

Rather than be treated as a 'bug' in the system, this could be billed as a 'feature': consumers can be given the option to select their own weights to create a personalized measure of educational quality. In fact, there are already examples of alternative ranking systems, such as *The New York Times'* "Build Your Own College Rankings," which publishes separate measures of institutional dimensions of relevance to those interested in attending. Though it does not generate the 'horse race' excitement of list of the 100 "best" colleges or hospitals, the personalization would potentially be more valid and useful to the end user and

allow the user to explore firsthand whether differences in weighting criteria affect their rankings. Instead of trying to describe a construct of abstract educational quality based on the *average* consumer's priorities, ranking providers can measure the personal construct of realized preferences for *every* consumer.

6.6. Develop a reform agenda with stakeholders at the table

Finally, this report could serve as an input to an effort to build consensus among higher education experts and other stakeholders on common concerns and on the path forward to potential improvements to college rankings. There is an opportunity to improve the quality of information presented to students and their families for use in decision-making about their college options.

Funding Statement

This project was funded by the Vanderbilt University Office of the Chancellor.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529-546.

Altbach, P. G. (2015). Academic inbreeding: Local challenge, global problem. *Asia Pacific Education Review*, 16(3), 317-330.

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices.* University of South Florida.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology*, 53(1), 605-634.

Bollen, K. A., & Lennox, R. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, 110(2), 305-314.

Bowman, N. A., & Bastedo, M. N. (2013). *Anchoring effects in world university rankings: Exploring biases in reputation scores*. In Higher Education in the Global Age (pp. 271-287). Routledge.

Bradburn, N. M., Cartwright, N., & Fuller, J. (2017). A theory of measurement. *Measurement in medicine: Philosophical essays on assessment and evaluation*, 73-88.

Carr, H. (2023). The WSJ/College Pulse College Rankings: Measuring Outcomes, Not Inputs. *The Wall Street Journal*. https://www.wsj.com/us-news/education/wsj-college-pulse-college-rankings-methodology-f010fc11

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, 38(2), 269-277.

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. Taylor & Francis.

Edwards, J. R., & Bagozzi, R. P. (2000). On the Nature and Direction of Relationships between Constructs and Measures. *Psychological Methods*, 5(2), 155-174.

Elliott, M. N., Zaslavsky, A. M., Cleary, P. D. (2006). Are finite population corrections appropriate when profiling institutions? *Health Services and Outcomes Research Methodology*, 6,153-156.

Fuller, A., & Korn, M. (2023). Colleges Urged to Produce Better Information on How They Spend Money. *The Wall Street Journal.* https://www.wsj.com/us-news/education/colleges-urged-to-produce-better-information-on-how-they-spend-money-8dc3b549

Furr, M. (2011). Scale construction and psychometrics for social and personality psychology. *Scale Construction and Psychometrics for Social and Personality Psychology*, 1-160.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 159(3), 385-409.

Groves, R. M. (2005). *Survey errors and survey costs*. John Wiley & Sons.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (Vol. 561). John Wiley & Sons.

Guarino, C., Ridgeway, G., Chun, M., & Buddin, R. (2005). Latent variable analysis: A new approach to university ranking. *Higher Education in Europe*, 30(2), 147-165.

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

Iacobucci, D. (2013). A psychometric assessment of the Businessweek, US News & World Report, and Financial Times rankings of business schools' MBA programs. Journal of Marketing Education, 35(3), 204-Ber219.

Itzkowitz, M. (2020). Price-to-Earnings Premium: A New Way of Measuring Return on Investment in Higher Ed. *Third Way*. https://www.thirdway.org/report/price-to-earnings-premium-a-new-way-of-measuring-return-on-investment-in-higher-ed.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research*, 30(2), 199-218.

Kelchen, R (2023). Discovering Issues with IPEDS Completions Data. https://robertkelchen.com/2023/11/06/discovering-issues-with-ipeds-completions-data/

Kerlinger, F.N. (1986). *Foundations of Behavioral Research* (3rd ed.). Fort Worth: Holt Rinehart and Winston.

Kim, J. & Shim, W-j. (2019). What do rankings measure? The U.S. News rankings and experience at liberal arts colleges. *Review of Higher Education,* 42(3), 933-964.

Krücken, G. (2003). Learning the 'New, New Thing': On the role of path dependency in university structures. *Higher Education*, 46, 315-339.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2-3), 292-326.

Marginson, S. (2014). University rankings and social science. *European Journal of Education*, 49(1), 45-59.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.

Morse, R. & Brooks, E. (2023). How U.S. News Calculated the 2019 Best Colleges Rankings. *U.S. News & World Report*. https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings

National Center for Education Statistics. (2023). About IPEDS. https://nces.ed.gov/ipeds/about-ipeds.

Nunnally, J. C. (1975). Psychometric theory—25 years ago and now. *Educational Researcher*, 4(10), 7-21.

Petter, S., Straub, D., & Rai, A. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, 31(4), 623-656.

QS Quacquarelli Symonds Limited. (2015). QS World University Rankings: Methodology. *QS World University Rankings*.

Ross, D. (2023). World University Rankings 2024 Methodology. *Times Higher Education (THE).*

Rounds, E. (2023). 2023 Price-to-Earnings Premium for Four-Year Colleges. *Third* Way. https://www.thirdway.org/report/2023-price-to-earnings-premium-for-four-year-colleges

Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, 74(1), 63-82.

Schreyögg, G., & Sydow, J. (Eds.). (2009). *The hidden dynamics of path dependence: Institutions and organizations*. Springer.

Shwartz, M., Restuccia, J. D., & Rosen, A. K. (2015). Composite Measures of Health Care Provider Performance: A Description of Approaches. *The Milbank Quarterly*, 93(4), 788–825. https://doi.org/10.1111/1468-0009.12165

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*. 15 (2): 201–292.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Stewart, J. (2013). New Metric for Colleges: Graduates' Salaries. *New York Times*. https://www.nytimes.com/2013/09/14/business/economy/nice-college-but-whatll-i-make-when-i-graduate.html

Tancredi, D. J., Bertakis, K. D., Jerant, A. (2013). Short-term Stability and Spread of the U.S. News & World Report Primary Care Medical School Rankings. *Academic Medicine*, 88(8), 1107-1115.

Whitehurst, G. J. & Chingos, M. M., & (2015). Deconstructing and reconstructing the college scorecard. *Brookings: Evidence Speaks Report*, *1*(4).

Whitford, E. (2023). How We Rank America's Best Colleges. *Forbes*. https://www.forbes.com/sites/emmawhitford/2023/08/29/how-we-rank-americas-best-colleges/

Appendix

Appendix 1. Additional Details about Five Exemplar College Ranking Systems

Table A1. Eligibility and Exclusion Criteria in the Five Exemplar College Ranking Systems

System	Eligibility Criteria (in 2024)	Exclusion Criteria (in 2024)
USNWR[2]	Four-year bachelor's degree-granting institutions included in Carnegie's Basic classification.	Excludes 'highly specialized' schools, >=100 undergraduates
WSJ [3]	All U.S. colleges that are Title IV eligible (accredited and eligible for federal financial aid), awards four-year bachelor's degrees.	Excludes insolvent institutions, service academies, institutions with <900 students. Government data for factors must be collected and made public and the accompanying student survey must receive >= 50 valid responses to the student survey (based on the recruits from two- and four-year college and universities (n>1500)). *University of Richmond was an example of a college that was excluded due to low survey participation (less than 50 student surveys) in 2024 despite ranking 63 in 2022.
Forbes [4]	Doctoral research, master's universities/colleges, and colleges with specialized programs in engineering/business/art based on Carnegie Classification (n≈1,481).	Excludes eligible universities announcing they are closing, fewer than 300 undergraduates, the five federal military academies (the last on the basis that they 'operate very differently from the other institutions').
THE[5]	All universities with a representative that self-submits through their data portal; university must teach at an undergraduate level.	Must meet some threshold of relevant publications in more than one 'narrow subject', must not be post-graduate only, must not have more than 80% of their publication output from one subject only, must have supplied "overall" numbers for the ranking year, must not have missingness in

[2] This report refers to the most recent version of the USNWR rankings at the time of writing, published in September 2023. As of June 14, 2024 the methodology for the 2024 USNWR rankings is publicly reported at the following URL: https://www.usnews.com/media/ai/education/2024BC-methodology .

[3] This report refers to the most recent version of the WSJ rankings at the time of writing, published in September 2023. As of June 14, 2024, methodology for the 2024 WSJ rankings have been publicly reported at the following URLs: https://www.wsj.com/us-news/education/wsj-college-pulse-college-rankings-methodology-f010fc11 (summary) and https://www.wsj.com/rankings/college-rankings/best-colleges-2024 (details).

[4] This report refers to the most recent version of the Forbes rankings at the time of writing, published in August 2023. Although it is referred to as the 2023 Forbes rankings, for consistency with the titles of other ranking systems, we refer to these as the 2024 Forbes rankings. As of June 14, 2024, methodology for the 2023 Forbes rankings are publicly reported at the following URL: https://www.forbes.com/sites/emmawhitford/2023/08/29/how-we-rank-americas-best-colleges/ .

[5] This report refers to the most recent version of the THE rankings at the time of writing, published in September 2023. As of June 14, 2024, methodology for the 2023 THE rankings are publicly reported at the following URL: https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2024-methodology .

System	Eligibility Criteria (in 2024)	Exclusion Criteria (in 2024)
		more than two critical values, must mark at least one subject as applicable, and must not be on their custom exclusions list defined as 'institutions that have requested not to participate in the ranking or that are not eligible for other institution-specific reasons' [3]
QS[6]	Though not explicitly stated, QS's universe appears to be all universities who submit data rather than some external frame like the Carnegie Classification. Within this group there are criterion such as at least three graduating cohorts, the granting of full undergrad and graduate degrees in at least two 'faculty' areas and two 'narrow subjects', and in-person teaching (see full eligibility criteria here).	Exclude eligible universities if they do not pass a reputation threshold (in the top 20% of academic reputation), research threshold (must have >=100 relevant papers indexed by Scopus), and size threshold (must achieve some other thresholds if the institution is less than 5000 students, see more details here).

Table A2: Components of Five Exemplar College Ranking Systems Displayed to Consumer

System	Components displayed (in 2024)
USNWR	• Primary scale • Summaries of some subcategory measures (e.g., tuition)
WSJ	• Primary scale • Secondary scale (0-100) • Category measures (scores) (e.g., student outcomes) • Summaries of some subcategory measures (e.g., average net price)
Forbes	• Primary scale • Summaries of some subcategory measures (e.g., tuition)
QS	• Primary scale • Secondary scale • Category measures (scores) (e.g., academic reputation) • Summaries of some subcategory measures (e.g., net tuition)
THE	• Primary scale

[6] This report refers to the most recent version of the QS rankings at the time of writing, published in June 2023. An announcement had been made for the 2025 rankings, however we were unable to access the rankings or any methodology reports. As of June 14, 2024, methodology for the 2023 QS rankings are publicly reported at the following URL: https://www.topuniversities.com/qs-world-university-rankings/methodology .

System	Components displayed (in 2024)
	• Secondary scale (0-100) • Category measures (scores) (e.g., teaching) • Summaries of some subcategory measures (e.g., F/M student ratio)

Table A3: Category Weights in Five Exemplar College Ranking Systems (2024)

System	Category	Weight	Justifications
USNWR[4]	*Graduation rates*	16%	Consumer preferences: UNSWR notes the top reasons why students attend colleges, "according to several different surveys ... relate to academic reputation, cost of attending and return on investment," which the methodology explicitly takes into account. Conceptual emphasis: Weight for Pell graduation rates was increased from 2.5% in previous rounds to 3% to "put more emphasis on a school's success supporting students from different socioeconomic backgrounds". Scope expansion or contraction: Weight for full-time faculty salary were decreased from 6 in previous rounds to 7% on the basis that 'faculty' was widened to include instructors, lecturers, and no-rank professors. Increasing source confidence: Weight for borrower debt increased from 3% in previous rounds to 5% based on increased confidence in the data source (NSLDS instead of financial aid surveys). Relative source confidence: Weight for first-generation graduation rates is lower than for Pell graduation rates because "the data is a year older and does not incorporate the entire bachelor's degree-seeking study." Compensation for category inclusion/removal: Weight for student-faculty ratio increased from 1% in previous rounds to 3% to compensate for the removal of class size; weight for graduation rate decreased from 17.6% in previous rounds to 16% to "make room for new factors" that incorporate similar information.
	First-year retention rates	5%	
	Graduation rate performance	10%	
	Pell graduation rates	3%	
	Pell graduation performance	3%	
	First generation graduation rates	2.5%	
	First generation graduation rate performance	2.5%	
	Borrower debt	5%	
	College grads earning more than a high school grad	5%	
	Peer assessment	20%	
	Faculty salaries	6%	
	Student-faculty ratio	3%	
	Full-time faculty	2%	
	Financial resources per student	8%	
	Standardized tests	5%	
	Citations per publication	1.25%	
	Field weighted citation impact	1.25%	
	Publications cited in top 5% of journals	1%	

System	Category	Weight	Justifications
	Publications cited in top 25% of journals	0.5%	**Missing values:** If <50% of ACT/SAT scores were reported by students to an institution, the standardized test score category was dropped and the weight for graduation rate is increased from 16% to 21% (on the basis that it is most strongly correlates with standardized test scores).
WSJ	*Student outcomes*	70%	**Conceptual emphasis:** For example, WSJ writes that "they now put greater emphasis on measuring the value added by colleges—not simply measuring their students' success, but focusing on the contribution the college makes to that success." **Fairness:** Input as a category is removed altogether; for example WSJ doesn't "factor in selectivity, which [they] consider to be an input, rather than an outcome for which the college should be *rewarded*."
	Learning environment	20%	
	Diversity	10%	
Forbes	*Alumni salary*	20%	**Product differentiation:** Forbes cites the Forbes American Leaders List in its rankings which "aims to gauge the leadership and entrepreneurial success of a college's graduates" as "part of what sets Forbes's rankings apart from the crowd." *(No other explicit justification could be found for either the categorization or the choice of weights.)*
	Debt	15%	
	Graduation rate	15%	
	Forbes American Leaders List	15%	
	Return on investment	15%	
	Retention rate	10%	
	Academic success	10%	
QS	*Academic Reputation*	30%	**Consumer preferences:** QS notes that it "significantly evolved [their] methodology" to reflect the "changing priorities of students"; This was the basis for the introduction of the sustainability and international research network categories in this round. **Expert judgment:** QS also notes the "collective intelligence" of the sector as a driver of its methodological refinement. This was the basis for an increase in the weight of employer reputation from 10% in the previous round to 15%. **Product differentiation:** QS writes about the usage of the employer reputation, for example: "we remain the only major ranking to focus on this vital aspect of a student's educational journey."
	Employer Reputation	15%	
	Faculty Student Ratio	10%	
	Citations per Faculty	20%	
	International Faculty Ratio	5%	
	International Student Ratio	5%	
	International Research Network	5%	
	Employment Outcomes	5%	
	Sustainability	5%	
	Teaching	29.5%	

System	Category	Weight	Justifications
THE	*Research Environment*	29.0%	**Content validity**: THE motivates the addition of five new indicators (13 to 18) with "most *comprehensive* and balanced comparisons". **Institutional diversity:** The THE methodology report says: "The 20th edition of the World University Rankings now sees another significant update to the methodology, so that it continues to reflect the outputs of the *diverse* range of research-intensive universities across the world, now and in the future."
	Research Quality	30.0%	
	Industry	4.0%	
	International Outlook	7.5%	

Table A4: Proxy Measures in the Five Exemplar Systems

System	Proxy measurements and data sources (not exhaustive)
USNWR	• *Social mobility* is measured via first-gen graduation rates (sourced via US Dept of Ed College Scorecard) and Pell grant students; also by % of grads from that college earned more than median workers in the same age cohort (0-100% score). • *Faculty resources* (and thereby student access to quality instruction) are proxied by average region-adjusted faculty salary from 2022 (sourced from AAUP), student-faculty ratio, number of full-time faculty. • *Faculty research productivity* is proxied via bibliometric data from Elsevier (e.g., citations per publication, citation in top journals). • *Borrower debt* is proxied by federal loan debt for borrowers who graduate. • *Standardized test scores* are proxied (if <50% of a college's class SAT/ACT score are reported) by graduation rates. • *First-gen performance/graduation measures* are proxied (if missing or if university had few federal loan recipients) via equivalent measures of Pell grant students.
WSJ	• *Value added* is proxied using median salaries by individual demographic group. • *Ethnic diversity* is proxied using the Gini-Simpson index of reported student demographic profiles.
Forbes	• *Academic success* is proxied via (1) the number of graduates who win Fulbright/Truman/Goldwater/Rhodes scholarships and (2) the average number of PhDs from that institution).
THE	• *Academic reputation* is proxied by perceptions of academic excellence via surveys of academics.
	• *Sustainability* is proxied via an aggregation of (1) M/F student ratio, (2) membership in sustainability associations, (3) views of staff on climate change

System	Proxy measurements and data sources (not exhaustive)
QS	commitment, and (4) public availability of sustainability strategies/commitments).

Table A6. Reported Normalization and Standardization Decisions in the Five Exemplar College Ranking Systems

System	Normalization and Standardization Decisions
USNWR	For each ranking (top liberal arts, top universities, top Northeast colleges, etc.) the 2024 USNWR transforms individual category scores into z-scores before weighting them; the overall score is then re-scaled so that the highest score falls at 100 (max-min rescaling). USNWR 2024 performed a different normalization of spending per student than previous years: "The spending per-student amounts were converted to percentile distributions before being normalized, which was a change from U.S. News' prior practice of taking the natural log of the data. Both had the impact of assuming diminishing returns on increased spending, but the percentile distribution approach further smoothed dispersed outlier values." USNWR 2024 uses both median (because it is "less impacted by outliers") and mean in creating a measure of borrower debt: "U.S. News averaged the median values from the two most recent College Scorecard releases, which reflected pooled cohort data from fiscal years 2019-2020 and 2020-2021. Averages for each school were calculated among all years for which data was available, which in some cases was only one year. UNSWR 2024 multiples categories of financial resources (e.g., total expenditures on instruction, research, etc.) by % of students attending who are full-time equivalent undergrads: • This requires an additional step of coding part-time undergrads and graduate students as 1/3 of full-time undergrad students. • Additionally, this spending per-student was converted to a percentile distribution and then normalized (a change from taking the natural log) which "had the impact of assuming diminishing returns on increased spending, but the percentile distribution approach further smoothed dispersed outlier values." USNWR 2024 annualizes salaries and then divides by the total faculty to produce an average salary for each school – this normalization is potentially problematic since many full-time faculty are on a 9 month salary and rely on grant funding during the summer period. Salaries are also normalized by cost-of-living / region (by CBSA or state if the school is outside of a CBSA). USNWR 2024 normalizes the ratio of Pell Grantee graduation rates to non-PG graduation rates to the 0-1 range (so if Pell Grantees overperform non-PGs, this is not reflected in the data). USNWR 2024 combines/standardizes ACT and SAT scores together in a series of steps: • SAT and ACT median scores across enrollees in each university is converted to 0-100 national percentile distributions. • Scores are weighted based on proportion of new entrants taking the test.

System	Normalization and Standardization Decisions
WSJ	All scores in the 2024 WSJ are normalized to the 0-100 scale before combining to form the overall score (see note at bottom). (Note: the 'years to pay off net price' is the only score in years/months and the methodology does not specify how this is normalized)
THE	The 2024 THE ranking standardizes each indicator via a cumulative probability function (a type of z-scoring).
QS	The 2024 QS ranking methodology makes use of normalization in various instances within category, however there is no mention of the exact normalization/standardization in the combining methodology; given that scores are out of 100, individual category scores are presumably normalized to the 0-100 scale. QS 2024's sustainability score incorporates a normalized version of race-to-zero commitment normalizing target year to a 5-50 score (e.g., before 2022 is given 50, 2026-2030 is 40, after 2060 is given 5); unclear why the floor is 5, not 0. QS 2024 applies a log transformation in the graduate employment index used to measure employment outcomes -- "we apply log-transformation 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."

Table A7. Truncation Decisions in the Five Exemplar College Ranking Systems

System	Truncation Decisions
USNWR	The 2024 USNWR only displays the ranks for the top 90% of universities and the decile range for the rest. This changed from top 75% in previous years.
WSJ	The 2024 WSJ ranking only displays institutions in the top 400 of scores (from their initial sample of eligible institutions; see below). This changed from 600 in the previous year.
Forbes	The 2024 Forbes ranking only displays institutions in the top 500 of scores (from their initial sample of eligible institutions; see below).
THE	The 2024 THE ranking displays rankings for the top 200 institutions by overall score; like USNWR, banded scores (by 50 rather than decile) are displayed ranging from [201, 250] to [1501+].
QS	The 2024 QS ranking does not describe any truncation, but it appears to be truncated at [1401+], [1201, 1400], [1001-1200], [951-1000], in 50's until 800 which goes in 10's until 600 (there are no scores for universities in these bands, only 'n/a'); after that certain rankings are skipped but 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.

Appendix 2: Total Survey Error Framework

The **total survey error** is the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data. The Total Survey Error framework focuses on the concept of total survey quality, where the key lies in the optimal allocation of resources to reduce the errors that may occur in survey (Groves et al., 2009). Therefore, a study design should consider all known sources of error, quantify them and the efforts are needed to focus on minimizing them to the extent possible.

It is worth noting that this framework defines the terms 'measurement' and 'representation' differently than the central Bradburn, Cartwright, and Fuller (2017) framework of the report. What distinguishes this framework from the last, is the identification of specific types of errors *as they occur in survey research*.

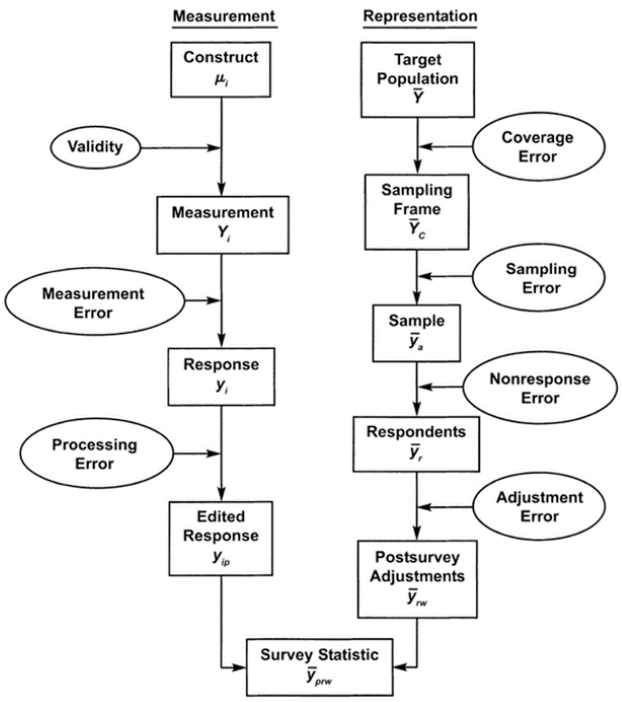


Figure A1. Survey Life Cycle from a Quality Perspective (Groves et al., 2009).

We now briefly highlight instances of each error category in the College Pulse survey, a prominent survey used in the construction of the WSJ rankings.

A.2.1. Validity

In the context of the TSE framework, validity may be understood as 'construct validity' as it is discussed in Section 2 of the main text.

The College Pulse survey, for example, is used to capture "a holistic view of student experiences and satisfaction", such as students' perception of learning opportunities, career preparation, dining halls, and sports facilities, and students' thoughts on diversity. Goldstein & Spiegelhalter (1996) argue that to ensure the validity of indicator variables as measures of effectiveness, a proper contextualization of outcome indicators is needed by taking account of institutional circumstances and the appropriate

specification of a statistical model. There is a challenge in understanding the integrity and appropriateness of data when used to compare institutional performance because student experiences may be more complex, involving spatial and prior experiences before college.

In short, the considerations laid out in the construct validity framework still apply to constructs captured by survey responses.

A.2.2. Measurement Error

Measurement error bias (also called systematic error) and measurement error variance (also called stochastic or random error) are two sources of measurement error in surveys. Measurement error bias can result from systematic deviation in responses from the true value due to poor measurement, such as misreading or misunderstanding a survey question. On the other hand, measurement error variance can result from variability across measurements.

College Pulse states that all surveys undergo question design and methodology review, and unreliable responders (such as those speeding or straight-lining) are detected and removed from surveys. It is unclear, however, how unreliable responders are detected and what percentage of responders are removed from the surveys. To take a concrete example of measurement error, USNWR 2024 peer assessment survey (presidents, provosts, and deans of admissions – or officials in equivalent positions) use 5-point scale to rate overall academic quality of peer schools' undergraduate academic programs from 1 (marginal) to 5 (distinguished). This scale, as with many custom-made Likert scales, may suffer from issues of *differential item functioning*, where individuals with different cultural contexts, demographic backgrounds, or cognitive states may interpret the same items on a scale differently (King et al., 2004). This may especially be a salient issue in the THE and QS reputation surveys that are administered globally. Fortunately, differential item functioning can be tested for and corrected with tools such as anchoring vignettes (King et al., 2004; King and Wand, 2007).

Measurement error in surveys can significantly impact the reliability of college rankings. Different cultural understandings of survey questions, variances in interpretation of terms like "research excellence" or "graduate employability," and the subjective nature of such assessments can contribute to measurement error. Additionally, the format in which questions are posed—whether allowing for open-ended responses or requiring selections from predefined options—can influence how responses are recorded and interpreted, potentially introducing bias (Groves et al., 2009).

A.2.3. Processing Error

Processing error can also arise from the transfer of data from the recording into the analytic dataset, such as data entry errors, misreading of respondent's writing.

College Pulse sends invitations to complete the survey to student email addresses and panelists obtain notifications in the smartphone app to take the survey. Students can complete the survey using the College Pulse App on iOS and Android platforms. They can also log in directly on the College Pulse website to complete the survey on their computers. Given that students can theoretically take the survey from multiple devices, it is unclear how multiple entries are handled and if there are any mode differences in responses (cellphone vs. computer) or differences (iOS vs. Android platforms) introduced by platforms that could cause any processing error or processing differences.

There can also be analytic error as well, when errors occur in the post-processing steps after the data collection, such as incorrect merging, attribution of response to the wrong individual, which are quite common in secondary analyses of survey data (West et al., 2016).

A.2.4. Coverage Error

Coverage error occurs when there is a gap between the target population and those represented in the sampling frame, which is the list of population members from which a sample is drawn. For example, if the sampling frame in the survey excludes specific individuals, such as those without cell phones, or students who live off-campus, the survey will suffer from coverage error.

College Pulse claims that all students must provide a `.edu` email address to join the panel and verify that they are currently enrolled either part-time or full-time in a four-year degree program. All invitations to complete surveys are sent using the student's .edu email address. This strategy presumes that all students actively use the .edu email address and that the survey invitation is not caught in the spam email. It could also miss transfer students or recent graduates who no longer use the `.edu` email address.

In the context of reputation surveys, coverage error emerges if the procedures for selecting survey participants do not account for all relevant academic and professional communities. For instance, newer or smaller institutions might be underrepresented if the sampling frame primarily consists of well-established universities. Similarly, reliance on digital platforms for distributing surveys can exclude potential respondents without reliable internet access, skewing perceptions towards more digitally connected populations (Groves et al., 2009).

A.2.5. Sampling Error

Another source of error can arise when there is a sampling error, which introduces variance in survey estimates. Sampling error can be reduced when we have a larger sample size. College Pulse states that it uses more than 20 demographic categories to target the most representative subsamples of the college panel based on the specific needs of each client and it requires every college included in the ranking received a minimum of 50 survey responses, with the majority receiving more than 100. It is unclear how the n=50 response requirement was determined and if different criteria apply to colleges of different enrollment sizes.

What is the ideal sample size for a survey? When using student survey responses to estimate college-level measures, several considerations need to be made to determine sample size. It is important to first determine the level of precision we would like to target using the surveys, such as the minimum effect size we would like to detect for comparing groups and the desired level of statistical power for any statistical testing. Based on this initial estimate of the sample size, further considerations will need to be made around the following questions:

1. Are the survey measures continuous or dichotomous? If continuous, what is the variability in the population? For dichotomous survey measures, the sample size calculation does not require prior knowledge of standard deviation of the population. However, if the measures are continuous, determining the sample size could be more complex and requires specification of the standard deviation of the population means.

2. Can we adjust for the missing data introduced by the nonresponse error? We can increase the sample size to adjust for missing data introduced by the nonresponse error. To do this, we need to know the nature of data missingness and whether one can assume data are *missing at random* under an assumed statistical model (Little & Rubin, 1989). See further discussions in Section 4.1.2.

What is the ideal sampling design? It is first important to consider whether the survey should use probability or non-probability sampling. If probability sampling is used, one would want to ask, "Does it use a simple random sample, cluster sample or stratified sample?" Each choice has different implications for sample size determination. If a non-probability sampling is used, one would want to ask if it is a convenience, snowball, or quota sample. It is NORC's view (and the view of much of the survey research community) that probability sampling, when feasible, is the ideal design given the known methodological issues with non-probability samples. Chief among these issues is that the lack of sampling frames and other information about the sampling process requires stronger analytic or weighting assumptions to achieve robust estimates of population quantities than equivalent surveys conducted via probability sampling.

The sampling design can influence the sample size needed to achieve the desired precision. For example, one might "oversample" one type of institution if a goal is to estimate performance by institution type. However, obtaining an estimate across all institutions would require adjusting for the fact that some institutions are over- (and other under-) represented in the sample, which would require a larger sample size to obtain comparably precise results to a simple random sample. Sampling error may also exist in surveys. For example, if a survey disproportionately reaches respondents from certain geographic regions or academic fields, their responses may skew the perception of institutional reputations in favor of those areas. In the context of global rankings like QS World University Rankings and THE College Rankings, which aim to capture worldwide perspectives, ensuring a balanced representation across countries, disciplines, and institution types is crucial yet challenging. The sampling strategy must carefully consider diversity to mitigate this error and ensure that the rankings reflect a genuinely global perspective (Groves et al., 2009).

A.2.6. Nonresponse Error

Another important source of error is nonresponse error, which occurs when respondents refuse to participate. Item nonresponse can also occur when certain items have missing value. These missing values can introduce bias to survey measures if they are *not missing at random* (Little & Rubin, 1989). Survey nonresponse can also result from outreach strategies, timing of the survey administration and method of participation and lack of participation may not indicate lack of student engagement. For example, University of Richmond was excluded due to low survey participation (less than 50 student surveys) in 2024 despite ranking 63 in 2022.

In some cases, our exemplars provide details on how nonresponse error is mitigated. The methodology report for the Times Higher Ed ranking system describes their imputation strategy:

> "Institutions provide and sign off their institutional data for use in the rankings. On the rare occasions when a particular data point at a subject level is not provided, we use an estimate calculated from the overall data point and any available subject-level data point. If a metric score cannot be calculated because of missing data points, it is imputed using a conservative estimate. By doing this, we avoid penalizing an institution too harshly with a "zero" value for data that it overlooks or does not provide, but we do not reward it for withholding them."

Here we see another example of fairness considerations (reward vs. penalty) guiding methodological choices as discussed in Section 3.2.

Nonresponse error is especially relevant in voluntary reputation surveys used for college rankings. High-status universities might be more inclined to respond, believing they have a vested interest in the outcome, while others may disregard the survey, assuming it has little impact on their standing or perceiving the exercise as irrelevant. The differential response rates can therefore bias the survey results towards the views of a particular segment of the higher education landscape, possibly those already enjoying greater visibility and prestige (Hazelkorn, 2015).

A.2.7. Adjustment Error

Postsurvey adjustments are usually made after a survey is conducted to reduce coverage, sampling and non-response error using known information about the target population, frame population or response rates of the sample. Adjustment error could result from the difference between the adjusted statistic and the population parameter due to inaccuracy in the known information about the target population.

College Pulse purports to apply "a post-stratification adjustment based on demographic distributions from multiple data sources, including the 2017 Current Population Survey (CPS), the 2016 National Postsecondary Student Aid Study (NPSAS), and the 2021-22 Integrated Postsecondary Education Data System (IPEDS) to rebalance the sample based on a number of important benchmark attributes, such as race, gender, class year, voter registration status, and financial aid status." It also uses an iterative proportional fitting (IFP) process to balance the distributions of all variables. College Pulse states that it trims weights to prevent individual interviews from having too much influence on the results, although it is unclear what kind of trimming strategy is used.

Appendix 3. Federal Committee on Statistical Methodology (FCSM) Data Quality Framework

In this section, we describe the data quality framework from the Federal Committee on Statistical Methodology (FCSM) that can be used to evaluate measures -- proxy measures in particular -- used in college ranking systems. We will show how such a framework offers a systematic approach and a common vocabulary to guide assessments of college ranking systems.

The FCSM Data Quality Framework was established by FCSM's Data Quality Analysis Working Group to provide practical information on identifying and reporting data quality for federal agencies and other organizations. Data quality is "the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust." The FCSM framework focuses on evaluating data quality in the context of statistical use and decision making. The FCSM Data Quality Framework (Figure A2) has three broad domains: utility, objectivity, and integrity, which are specific areas where data quality can be considered.

Utility refers to "the extent to which information is well-targeted to identified and anticipated needs; it reflects the usefulness of the information to the intended users." The utility domain covers five areas:

- *Relevance*, which refers to whether the data product is targeted to meet current and prospective user needs (achieved when the scope, coverage, reference period, geographic detail, data items, classifications, and statistical methodology meet user needs)

- *Accessibility*, which relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users.

- *Timeliness*, the length of time between the event of phenomenon the data describe and their availability

- *Punctuality*, measured as the time lag between the actual release of the data and the planned target date of data release

- *Granularity,* which refers to the amount of disaggregation available for key data elements and can be expressed in units of time, level of geographic detail available, or the amount of detail available.

The second domain, objectivity, refers to "whether information is accurate, reliable, and unbiased, and is presented in accurate, clear, interpretable, and unbiased manner." Under the objectivity domain, there are two areas for consideration:

- *Accuracy and reliability*. Accuracy refers to the closeness of an estimate from a data product to its true value, and reliability focuses on the consistency of results when the same phenomenon is measured more than once under similar conditions.

- *Coherence*, the second dimension under objectivity domain, is "the ability of the data product to maintain common definitions, classification, and comparability with other relevant data."

Finally, integrity refers to "the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision." The integrity domain encompasses four areas:

- *Scientific integrity*, which refers to "an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence"

- *Credibility*, which refers to the "confidence that users place in data products based simply on the qualification and past performance of the data producer"

- *Computer and physical security*, which refers to the "protection of information throughout the collection production, analysis and development process from unauthorized access or revision to ensure that information is not compromised through corruption or falsification," and

- *Confidentiality*, which refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party.

It is difficult to assess each dimension of the integrity domain for the data elements used in the college ranking systems with publicly available information alone. Therefore, in this report, we discuss all dimensions under the Utility and Objectivity domains but limit our discussion of the Integrity domain to

the *scientific integrity* and *credibility* dimensions. A complete discussion of the credibility of all data input for each data element may be beyond the scope of this report, but we will discuss them when principles are relevant.

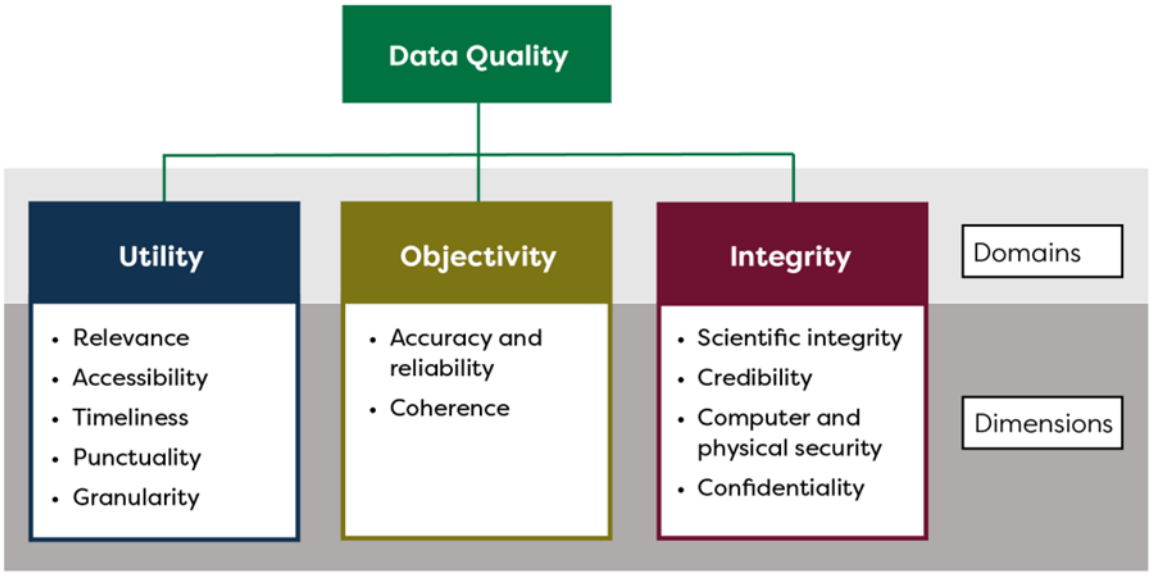


Figure A2. The FCSM Data Quality Framework (FCSM-20-04), 2020
(Federal Committee on Statistical Methodology: A Framework for Data Quality (FCSM-20-04))

Appendix 4. Berlin Principles on Ranking Higher Education Institutions

In recognition of the importance of higher education rankings for consumers, an international consortium founded by UNESCO published, in 2006, the Berlin Principles on Ranking Higher Education Institutions (or 'Berlin Principles'), a framework for the design and dissemination of higher-ed rankings.

The 16 principles fall under four thematic pillars and are reproduced from its original source below (noting that many, if not, most principles overlap with messages in this report):

A.4.1. Purposes and Goals of Rankings

1. Be one of a number of diverse approaches to the assessment of higher education inputs, processes, and outputs. Rankings can provide comparative information and improved understanding of higher education, but should not be the main method for assessing what higher education is and does. Rankings provide a market-based perspective that can complement the work of government, accrediting authorities, and independent review agencies.

2. Be clear about their purpose and their target groups. Rankings must be designed with due regard to their purpose. Indicators designed to meet a particular objective or to inform one target group may not be adequate for different purposes or target groups.

3. Recognize the diversity of institutions and take the different missions and goals of institutions into account. Quality measures for research-oriented institutions, for example, are quite different from those that are appropriate for institutions that provide broad access to underserved communities. Institutions that are being ranked and the experts that inform the ranking process should be consulted often.

4. Provide clarity about the range of information sources for rankings and the messages each source generates. The relevance of ranking results depends on the audiences receiving the information and the sources of that information (such as databases, students, professors, employers). Good practice would be to combine the different perspectives provided by those sources to get a more complete view of each higher education institution included in the ranking.

5. Specify the linguistic, cultural, economic, and historical contexts of the educational systems being ranked. International rankings should be aware of possible biases and be precise about their objective. Not all nations or systems share the same values and beliefs about what constitutes "quality" in tertiary institutions, and ranking systems should not be devised to force such comparisons.

A.4.2. Design and Weighting of Indicators

6. Be transparent regarding the methodology used for creating the rankings. The choice of methods used to prepare rankings should be clear and unambiguous. This transparency should include the calculation of indicators as well as the origin of data.

7. Choose indicators according to their relevance and validity. The choice of data should be grounded in recognition of the ability of each measure to represent quality and academic and institutional strengths, and not availability of data. Be clear about why measures were included and what they are meant to represent.

8. Measure outcomes in preference to inputs whenever possible. Data on inputs are relevant as they reflect the general condition of a given establishment and are more frequently available. Measures of outcomes provide a more accurate assessment of the standing and/or quality of a given institution or program, and compilers of rankings should ensure that an appropriate balance is achieved.

9. Make the weights assigned to different indicators (if used) prominent and limit changes to them. Changes in weights make it difficult for consumers to discern whether an institution's or program's status changed in the rankings due to an inherent difference or due to a methodological change.

A.4.3. Collection and Processing of Data

10. Pay due attention to ethical standards and the good practice recommendations articulated in these Principles. To assure the credibility of each ranking, those responsible for collecting and using data and undertaking on-site visits should be as objective and impartial as possible.

11. Use audited and verifiable data whenever possible. Such data have several advantages, including the fact that they have been accepted by institutions and that they are comparable and compatible across institutions.

12. Include data that are collected with proper procedures for scientific data collection. Data collected from an unrepresentative or skewed subset of students, faculty, or other parties may not accurately represent an institution or program and should be excluded.

13. Apply measures of quality assurance to ranking processes themselves. These processes should take note of the expertise that is being applied to evaluate institutions and use this knowledge to evaluate the ranking itself. Rankings should be learning systems continuously utilizing this expertise to develop methodology.

14. Apply organizational measures that enhance the credibility of rankings. These measures could include advisory or even supervisory bodies, preferably with some international participation.

A.4.4. Presentation of Ranking Results

15. Provide consumers with a clear understanding of all of the factors used to develop a ranking, and offer them a choice in how rankings are displayed. This way, the users of rankings would have a better understanding of the indicators that are used to rank institutions or programs. In addition, they should have some opportunity to make their own decisions about how these indicators should be weighted.

16. Be compiled in a way that eliminates or reduces errors in original data, and be organized and published in a way that errors and faults can be corrected. Institutions and the public should be informed about errors that have occurred.