Comparative analysis of graduate statistics department using perception based ranking and research based ranking

by

Cameron O'Neill

Bachelors of Science, Biology, Westmont College, 2014

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Sciences

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Cameron O'Neill

It was defended on

March 23rd 2023

and approved by

Lucas K. Mentch, Ph.D., Department of Statistics

Yu Cheng, Ph.D., Department of Statistics

Kehui Chen, Ph.D., Department of Statistics

Comparative analysis of graduate statistics department using perception based ranking and research based ranking

Cameron O'Neill, M.S.

University of Pittsburgh, 2023

University rankings have become an important part of the admissions processes and their audience has grown substantially over the last two decades. Yet there is still a lack of coverage for graduate programs outside of business, law and medical school. One of the eminent rankings is the U.S. News & World Report, which has begun ranking graduate programs through peer assessment survey data which relies heavily on institutional reputation. The purpose of the study was to estimate the difference between a perception-based ranking measured through the USNWR rankings and a research-based ranking measured through Google Scholar metrics for 96 graduate statistics department within the United States. Additionally, the researcher sought to examine the underlying bias that might be present in the two rankings. A research rank statistic was generated for each department based off the mean of the normalized Google Scholar metrics for each professor within their department. Lastly, a bootstrap sample was conducted and was used to calculate the possible ranges of values departments could take. This was used to determine if a department deviates measurably between its research rank and perception-based rank. The results showed there is a measurable difference between the perception-based ranking and the research-based ranking.

Table of Contents

List of Tables

List of Figures

Preface

When I first started the process of this thesis, I was nervous and had many doubts. But throughout the process my advisor Dr. Lucas Mentch and my family helped support me and are the reasons I was able to finish. Without their patience, and wise advice I wouldn't been here today.

The other committee members, Dr. Yu Cheng and Dr. Kehui Chen I will be forever grateful, not only for the help they gave me throughout the year, but also for their kindness in their comments and review.

Lastly, I want to give a huge thanks to the fellow students that helped make this project possible, Marc Richards and Junyi Bai. Without their work and help this project wouldn't have been possible their collaboration on the project was vital.

1.0 Introduction

Global higher education rankings have become a critical tool for students, faculty, departments, universities and governments to track academic progress and standing compared to their global and relative peers. From 1950 to 2010 [Barro and Lee, 2013] Barro Lee [2010] and [Lee and Lee, 2016] Lee and Lee (2016) observed a steady increase in mean school years across the globe. The increase has lead to the development and establishment of more universities and colleges. Based off the latest estimates as of 2022, there are approximately 25,000 universities and colleges worldwide.

The rise of mean schooling years has lead to the increase in the number of university students across the globe. The sudden growth has seen a dramatic increase in higher education rankings spanning across different regions, languages and fields of study to help provide information to the growing amount of university students looking to study domestically and internationally.

With such great demand, a system for ranking or measuring universities based off different criteria emerged. The United States in particular has a long history of ranking universities, with one of the first published lists dating back to 1906 and another updated list in 1910 [Myers and Robe, 2009]. The main objective of these early lists until the 1950s and 1960s were based solely on ranking off the number of eminent alumni scientists or faculty produced by the universities, which acted as an outcome based ranking [Myers and Robe, 2009]. It was during this time period when U.S. News and World Report (USWNR) published their first reputational survey university ranking in 1983. The USWNR ranking saw immediate success and slowly became an institution in and of itself. Additionally, in 1982 the National Academy of Sciences conducted one of the first of its kind studies examining approximately 3,000 doctorate granting programs under both non-reputational and reputational survey metrics.

Over the course of the following decades rankings, specifically the USNWR rankings, have become extremely important to applicants and admissions departments [Meredith, 2004]. Yet at the same time, research dating as far back as 1998 has cast doubts on the rankings themselves [Machung, 1998]. There has also been a growing number of rankings with their own methodology and criteria with often times contradictory results such as Academic Ranking of World Universities (ARWU), U.S. News and World Report (USNWR), The Times of Higher Education (THE) and the QS World University (QS) rankings to name a few.

The USNWR rankings have become one of the leading indicators and rankings for American universities. With the ubiquitous nature of the USNWR university rankings, additional research has been conducted to examine their methodologies and effectiveness. There is growing evidence that universities stay relatively stagnate from year to year [Dichev, 2001], and the USNWR may weight academic reputation more heavily than the data insists [Webster, 2001].

Additionally, the USNWR relies solely on survey data to rank specific graduate departments, which poses a unique challenge to unique and emerging fields, such as statistics, data science, machine learning and artificial intelligence that have unknown reputional metrics.

1.1 Statement of the Problem

Within the United States, the predominate and industry leading university rankings belong to the USNWR. These rankings overwhelming target undergraduate reputation and metrics. The National Academy of Sciences found there has been a steady increase in individuals seeking doctorates within the United States (The National Academy of Sciences). Due to this increase demand the USNWR began ranking graduate departments. The current USNWR graduate department rankings rely only on peer assessment self reported survey data. The graduate department peer assessment survey data collected by the USNWR can act as a proxy for measuring the perception of a graduate statistics department.

The USNWR has done an excellent job of filling this void for specific graduate studies, such as law, medicine and law. But when examining specific departmental graduate level rankings in STEM, the reported response rates for graduate department science programs by USNWR was biological sciences, 15.8; chemistry, 30.6; computer science, 41.1; earth sciences, 24.7; mathematics, 33.6; physics, 27.9; statistics, 46.6; and biostatistics, 59.1.

The current USNWR graduate statistic department rankings rely solely on survey data with a response rate of 46.6 percent resulting in a total of 101 universities being ranked. The survey is designed on a score scale of 1 (marginal) to 5 (outstanding) with the highest department receiving a score of 4.9 and the lowest departments receiving a score of 1.5.

Across disciplines, topics, and format, self reported survey data has been demonstrated to introduce bias at varying degrees [Kreuter et al., 2008]. Additionally, when survey data is relied to rank graduate departments, this method has a difficult time producing rankings for developing fields and disciplines such as statistics, machine learning, data science, and artificial intelligence where reputational data is sparse due to the relatively young age of these departments.

The National Academy of Sciences has attempted to ameliorate this issue through the agencies *A Data-Based Assessment of Research-Doctorate Programs in the United States* by collecting their own data and publishing the majority of it as well as their scoring metrics and standardized scores. But this data is obtuse to access, and is rarely updated more than every decade (it was last updated in 2011 but still relied on data collected in 2005 and 2006).

A multitude of methods have been used to examine and study the USNWR rankings specifically due to their legacy and impact on the admissions process and departmental reputation [Bowman and Bastedo, 2010]. Most notability the variation year to year appears to be mostly variance or ”noise” in the data [Clarke, 2002] and [Dichev, 2001]. Similarly, principal component analysis conducted on the USNWR rankings and concluded the 11 ranking criteria outlined by USNWR differed substantially from the weighting scheme [Webster, 2001].

The USNWR graduate department rankings have a measurable impact on admissions and organizational reputation [Bowman and Bastedo, 2010]. Yet one of the key indicators for graduate departments is the ability to generate novel and impactful research, which in turn creates additional funding and opportunities for graduate students. This study seeks to understand the difference of perception, measured through USNWR rankings, and a research output based ranking to provide students, faculty, and researchers an understanding of how divergent perception is from research output.

1.2 Purpose of the Study

The purpose of this study is to compare and contrast the USNWR graduate statistic department rankings to research output based rankings measured by professor information and their respective Google Scholar indicators and quantify USNWR deviance from research indicators.

Bastedo and Bowman established a well defined connection between USNWR ranking and organizational reputation and student admission/application decisions. Thus there is a clear motivation for determining a research oriented ranking that will provide a different context and less bias through a non-survey based approach to ranking graduate departments.

The current USNWR graduate statistic department rankings rely solely on survey data with a response rate of 46.6 percent resulting in a total of 101 universities being ranked. The survey is designed on a scale of 1 (marginal) to 5 (outstanding) with the highest department receiving a score of 4.9 and the lowest departments receiving a score of 1.5. The scores are then averaged and ranked numerically from highest to lowest. Regardless of previous studies showing a high level of noise [Clarke, 2002] and [Dichev, 2001], the data would be prone to high levels of noise due to social desirability bias, response rate bias and order-effect bias.

Research output data has never been easier to collect between Google Scholar, ResearchGate, Web of Science, and personal websites. It seems logical to examine a research output based ranking examining key research indicators such as total cites, i10 index, and h-index. These three citation and research metrics provide a well rounded approach. Total cites measures the total number of citations across all publications on which the faculty member is an author. H-index is defined by Google Scholar as ”the largest number h such that at least h articles in that publication were cited at least h times each”. i10-index meaures the number of papers with at least 10 citations. These metrics are widely accepted and used consistently across departments and discipline.

We collected data on 96 of the 101 statistics graduate departments listed in the USNWR rankings. The information for each university includes only tenure professors within the statistic department and those that had Google Scholar profiles which resulted in a dataset consisting of 1460 professors. For each of the 1460 professors, university, title, PhD year, PhD granting institution, start year, total cites, i10-index and h-index was collected.

1.3 Research Questions

The main aims of the research conducted can be broken down into the following questions.

- Is there a measurable deviance between perception of graduate statistic department, measured through the USNWR ranking and a research output based ranking measured through Google Scholar metrics?
- What type of bias, if any, is present in the perception based rankings for graduate statistics departments when compared to a research output based ranking?
- Is there any discernible pattern in the schools that benefit from a research output based ranking versus perception based and vice versa?
- What role does size of the department play in assessing the perception of a statistics department?

1.4 Ethical Considerations

The study required using publicly available professor information such as names, titles, PhD granting institutions, start year and PhD year. The information collected was all publicly accessible from departmental web pages or personal websites. If information was not easily accessible there was no attempt to try and find this information through other non-publicly accessible means.

2.0 Literature Review

University rankings have become an important part of higher education due to their ubiquity and impact on higher education. More specifically, students want to understand the ranking of universities and specific departments to make an informed decision on where to apply and accept offers. Faculty and universities want to be able to measure themselves against their peer universities, estimate the impact of administrative initiatives, and how individual faculty members research output compares to their peers. Currently, most rankings for graduate departments outside of medicine, law and business rely on perception of graduate departments measured through self reported peer assessment surveys.

We examined the difference between a perception based ranking measured through the USNWR rankings and a research output based ranking measured through Google Scholar metrics for the 101 graduate statistics departments ranked by the USNWR. Additionally, we examined the bias that both types of rankings confer and which factors could be contributing to those biases.

2.1 Impact of University Rankings

In 2001 the U.S. News & World Report estimated that it sold over 2.2 million copies of the college rankings issue, reaching an audience of nearly 11 million people [Dichev, 2001]. Yet up until that point there had been little to no empirical analysis of the rankings impact. It was in 1999 that Monks and Ehrenberg conducted one of the first of its kind studies examining the impact of the USNWR college rankings [Monks and Ehrenberg, 1999].

The purpose of the study by Monks and Ehrenberg in 1999 was to conduct an empirical analysis of the impact of rankings on applications, admission, enrollment decisions and institutions tuition rates. Additionally, the researchers wanted to try and understand the interplay between the variables and how the ranking increasing or decreasing maybe responded too by the university or college.

The two researchers examined admit rates, yield rates, average SAT score, freshman class size and tuition change for 16 of the top 25 national universities and 13 of the top 25 liberal arts colleges according to USNWR. The data from these universities and colleges are from the 1988 to 1999 academic years. In total there are 330 observations for 30 universities or colleges across 11 years.

The impact of the rankings was measured through the lagged USNWR ranking when regressed on the yearly statistics stated above. To help control for differences across the varying universities and colleges the average endowment per student was used. Additionally, there were numerous binary variables such as, fell out of top 25, institution, and year effects.

The impact of the USNWR rank on admissions outcomes were surprisingly moderate. An increase or decreasing by one rankings spot resulted in a statistically significant increase of 0.399% in the admission rate. This same general pattern held for yield rate as well, but with a smaller magnitude. Similarly, the average SAT was statistically significant for rank, which went down by three points each rank increased (where increasing rank is moving from smaller to larger rank values). Most notably increasing the the rank by 10 places caused aid adjusted tuition to decrease by 4%.

The analysis from this study indicates that universities and colleges in their sample appeared to be responding or at least being impacted in measurable ways by the latest USNWR rankings. Most notably it is impacting both the external perception, such as average SAT and internal perception, increasing financial aid, or reduction in aid adjusted tuition costs.

A major limitation of the study was the lack of accounting for methodological changes in weights by the USNWR that produced the rankings every year. This could account for some of the changes and movement within the rankings and thus reduce the significance of the variables measured. Additionally, the sample they chose, Top 25 in the two categories, is a limitation as well. These universities and colleges are competing for the strongest high school students, and it seems logical these students would care most about the latest rankings because they are hyper engaged in the process. Thus if the researchers examined say the Top 100 in each category, this trend may disappear.

In 2004 Marc Meredith sought to expand upon the work that Monks and Ehrenberg conducted in 1999 by expanding the number of universities and colleges present in an empirical analysis of the USNWR rankings [Meredith, 2004]. Meredith also included demographic information to try and ascertain how socioeconomic and racial demographics played a role.

The dataset for Meredith's study used USNWR Best colleges issues from 1991 to 2000 for 233 universities and colleges. The data collected in those issues were average SAT scores, acceptance rates, and the percentage of students in the top 10% of their class. Rank was classified as being inside the Top 25 and outside of the Top 25. In addition the researcher collected Princeton Review college tuition data.

The demographic information was collected from the Integrated Postsecondary Education Data System, which included ethnicity enrollment information and the value of gifts/grants/contracts/Pell Grants. One drawback was that ethnicity data was only available for 1990 to 1998 and the financial data was only available from 1990 to 1996.

The methods for the study differed slightly from Monks and Ehrenberg by including two models, one to model each outcome, such as average SAT score, acceptance rate and percent in the top 10% of high school class, with a rank and year term with a random shock term. The next model used the outcomes generated from the first model to build a fixed effect model with an assumption by the researcher that universities have an invariant unobserved effect on admission outcomes.

The predictors in Meredith's study included the rank variable, a top 25 binary variable, and four quartile dummy variables to indicate weather the school is ranked in the 1st, 2nd, 3rd, or 4th quartile.

The first model which examined specific outcomes from USNWR rankings showed that movement between the 1st quartile and the 2nd quartile had a significant impact on admissions statistics. Specifically, the percent of top 10% high school applicants went up by 1.5% and the acceptance rate decreased by 4%. These effects were the same for each increase in quartile, but the lower the starting quartile, the less gains were made. For example, going from the 3rd quartile to the 2nd quartile saw an increase in top 10% high school students by only 1.4% and a decrease in acceptance rate by 1.0%.

The conclusions from the study confirm the previous work done by Monks and Ehrenberg that being inside the first quartile dramatically increases your perception. At the time of the study the primary medium for reading about the rankings was in the physical magazine. Meredith hypothesized that being in the 1st quartile puts you on the same page as the Top 25 schools thus, there is an order effect bias at play.

Similar to Monks and Ehrenberg, Meredith's study could be greatly enhanced by incorporating a variable or metric that accounted for small changes in methodology on the part of USNWR.

In previous studies, as mentioned above, the USNWR had been shown to impact student preference, but there has been less research trying to understand its effect on organizational, or institutional standing. It stands to reason that if USNWR impacted the admission process and student preference it could also impact the perception of organizational performance and standing.

A study conducted by Bastedo and Bowman in 2010 sought to examine the effect the USNWR undergraduate ranking has on the peer assessment survey's that are included as part of the USNWR yearly undergraduate rankings [Bowman and Bastedo, 2010]. Specifically, the researchers examined if controlling for changes in educational quality and financial information do college rankings influence future peer assessments. Additionally, examine USNWR Tier levels, specifically, if controlling for the same things, what impact does moving between the four USNWR Tiers have on peer assessment ratings.

The study used data from 1989, 1995, 2000 and 2006 from the USNWR Top 25 national universities and the Top 25 national liberal arts colleges. The data consisted of each years overall ranking and the peer assessment rating for the year. Four control measures were added for each year, graduation and retention rank, faculty resources rank, selectivity rank and financial resources rank. Additionally, they examined data from 1995, 2000, and 2006 with 168 national universities and 119 liberal arts colleges, that were labeled as tier 1, tier 2, tier 3, or tier 4 in addition to the previous variables.

The researchers used a structural equation model (SEM), which was chosen to address the high multicollinearity established within the metrics. Using the SEM the covariance matrices were examined using maximum likelihood estimation. The two sets of universities were analyzed separately and then both together. The years included were 1989-95, 1995-2000, and 2000-2006. The same process was repeated with tier labeled universities, except tier was dummy coded.

The results showed that for both liberal arts colleges and national universities the 1989 USNWR rankings predicted the 2006 peer assessment ratings at the 0.05 statistical significance level. Additionally, when all the institutions were included, the effect was still generally the same. The rankings in 1989 have a significant effect on the peer assessment ratings in all subsequent years. This general trend is held when examining 1995 rankings impact on 2000 peer assessment ratings and 1995 rankings on peer assessment ratings in 2005. It is important to note that the artifactual models did not predict vice versa, thus peer assessment scores from 1989 did not impact the 2005 overall rankings at a statistically significant level.

The researchers also examined the effect of moving between tiers while controlling for educational quality, and financial information and determined it had a significant effect for national universities and a weak effect for liberal arts colleges.

The results showed a clear indication that for both national universities and liberal arts colleges future peer assessment of reputation are impacted by ranking, tier level and changes in tier levels. More specifically, moving between tier 1 and tier 2 saw a significant change in peer assessment score. This highlighted one of the challenges of the USNWR ranking system that a university could move from tier 1 and tier 2, which is just one spot and see huge peer assessment reputation change. Additionally, the tier level only appeared to impact schools in tier 2, and tier 3. Lastly the study highlights the problem that rankings or tier's don't convey the true difference between schools. For example, the last tier 1 school and the first tier 2 school share more in common than the first tier 1 school and last tier 2 school. This same principal also applies to lists, where it is hard to understand the difference between say 9th and 10th. It could be that there is a large drop off between 9th and 10th, but a list conveys does not convey this.

The research literature indicates rankings can have a large impact on both student admission statistics, institutional decision makers and how other academics view other peer universities and departments. These findings from the three research papers above provide evidence for the need to take great care in building a methodology for ranking universities and colleges. Also, peer perception is impacted by rankings, thus there is a feedback loop when relying only on peer assessment survey's when ranking universities favors initial momentum in the ranking. The current study seeks to remedy this situation by relying on research based metrics.

2.2 Understanding University Rankings

Most of the research preceding 2000 was focused primarily on understanding the impacts the USNWR were having on the admissions process as it was still debated whether they were having a measurable impact on the college selection process. The studies discussed in the previous section established the connection between the rankings and admission decisions. The next group of studies then established the connection between internal university decision makers and USNWR rankings. There was a shift from measuring their impact to seeking to understand if the rankings were measuring what they attempting too.

Interested observers of the USNWR university rankings noticed that most schools were relatively stagnate within the rankings. With few exceptions, most schools remained in their quartiles from year to year. In 2001 Ilia Dichev attempted to estimate the noise in the USNWR rankings [Dichev, 2001]. The purpose of the study was to investigate the predictability of the rankings year to year changes and to test reasonable hypothesis proposed by the researcher.

The study used the Top 25 national universities as determined by the USWNR ranking and the Top 25 liberal arts colleges as determined by the USNWR ranking. The data was collected for the 50 schools over the course of 1987 to 1998. The total available observations was 243 for national universities and 236 for liberal arts colleges because some years schools would move inside and outside of the Top 25.

First, the change in year to year ranking was calculated for each university/college for each year. This was the main variable in the time series model. Upon examining the distribution of the changes in USNWR ranking and testing for normality assumptions the researcher decided to remove extreme values. A time series was then fit to the data for each group of universities testing up to four lags.

The results from fitting multiple time series models showed that a model with the first and second order lagged changes produced the best combination of explanatory power and parsimony. For both national universities and liberal arts colleges, the coefficients produced for the first and second order lagged terms were both negative, relatively large, and statistically significant at the 0.001 level.

In order to confirm the results produced in the time series models, the probability of obtaining a positive or negative ranking changes conditional on the signs of the lagged terms was calculated. The probability of a switch in ranking signs was 0.77 for national universities and 0.75 for liberal arts colleges. Next, the researcher attempted to estimate the noise in the ranking changes by setting up a system of equations using quantitative estimates of the variances, covariances and ranking change to estimate permanent and transitory changes. The results were that 65% of movement in the national university rankings will revert in the next ranking and 79% for liberal arts colleges.

The majority of year to year movement within the USNWR rankings appears to be mostly stable over longer time periods of examination but noisy from year to year. This seems reasonable because the core USNWR model is not fundamentally radically changing year to year, but slightly methodological changes are introduced year to yer. It also takes a great deal of momentum for the perception of a university to change and thus impact admissions over a long enough period to induce systemic decline or incline. The author suggests the idea of a "fight-back" effect and a "complacency" effect after a poor ranking or sustained success respectively. But when examining universities and colleges over a 5 year change, the first order term was not significant indicating no such impacts.

Perhaps the most compelling explanation is the underlying data in which the USNWR model is ingesting is extremely noisy. The author gives an example where John Hopkins University was 9 in faculty resources in 1989, 30 in 1990 and 9 again in 1991 followed by a 37, 97, 15 from 1993 to 1995. This impacted their overall ranking by going from 14, 15, 11, 15, 15, 22, and then 10. Dichev states "It is difficult to understand how a proper measure of the faculty resources of John Hopkins could ever manifest such a change. In fact, the magnitude and the pattern of these changes simply defy common sense about how true faculty resources might evolve at one of the most elite U.S. universities". One of the most challenging aspects of generating an aggregated ranking based off metrics from universities, is that you must rely on the university to self report the right information and that includes accidental human error.

One major limitation of the study is perhaps the over simplified view of rankings and what influences them. Controlling for number of applicants, economic factors based on year would have added important contextual information to the models that could have helped account for some of the noise encountered.

Additionally, the study had a relatively small and elite sample size. There is an argument, as put fourth by future research conducted by Meredith, that the 1st quartile is a different subset that is impacted more heavily than the vast majority of universities that occupy the other 3 quartiles. These schools are much more susceptible to the noise in the data. It seems reasonable to assume that elite schools mostly rotate around in the Top 25, while schools in the the lower 3 quartiles could be making substantial changes because they have more room for movement. For example with a recent example, Northeastern University moved from 96 to 42 from 2012 to 2019 and similarly, Texas Christian University moved from 108 to 76 within the same time frame. These are unlikely to be due to noise, and represent probably lasting changes.

Similar to the previous study Thomas Webster was primarily concerned with understanding the methodology and metrics being used by the USNWR to produce their tier rankings [Webster, 2001]. More specifically Webster was interested in understanding how each of the then 11 ranking criteria published by the USNWR under their methodology section actually impacted the tier rankings.

The purpose of the study was to use principle component analysis (PCA) to examine the 11 ranking criteria as outlined in the USWNR methodology, specifically, examine the relative contributions of each criteria and compare it to the USNWR weighting scheme.

The data used in the study was the 1999 USNWR College Rankings which included 228 national universities and 162 national liberal arts colleges. For the study, only national universities reporting SAT scores were used. The average SAT scores for each national university was calculated by using the USNWR first and fourth quartiles average SAT scores.

The other variables included in the PCA was: ACCRAT ratio of the number of students to the number of applications for admission, ACTGRAD the 6 year graduation rate for students, ALUM the percentage of undergraduate alumni of record who donated money to the institution, FTFAC the proportion of total fcaulty employed on a full time basis, LT20 the percentage of undergraduate classes with fewer than 20 students, MT50 the percentage of undergraduate classes with 50 students or more, PREDGRAD the predicted graduation rate, REP the average rating of the quality of a school's academic programs as evaluated by officials by similar institutions, RET the ratio of the number of students admitted to the number of applicants, TOP10 proportion of students enrolled in university who graduated in the top 10% of their high school.

Principal component analysis was conducted on the 11 ranking criteria outlined above. The results showed that the first two principle components accounted for 79% of the total variation. The first principal component showed roughly equal loadings on only 8 of the 11 rankings criteria. The eight ranking criteria in decreasing order of importance from the 1st principal component are SATAVG, PREDGRAD, ACTGRAD, REP, RET, ACCRAT, TOP10, ALUM.

In addition, the researcher also fitted an ordinary least squares regression with the 11 ranking criteria, which showed that only six of the 11 ranking criteria were statistically significant. Most notably, the signs of the coefficients with regards to ALUM, SATAVG and TOP10 seem to indicate the opposite relationship. The researcher examined the pairwise correlation coefficients and found evidence of substantial multicollinearity. This confirmed the need and reasoning behind conducting PCA.

The most striking result of the PCA analysis was that the most heavily weighted USNWR ranking criterion in 1999 academic reputation was ranked fourth in the 1st principal component. Similarly, the most important first principal component variable based of the eigenvalue was average SAT scores, which was only weighted 6% by the USNWR ranking.

As the author notes, the main limitation of the study is having to self select universities that have self reported SAT scores to USNWR. While the author acknowledges this could have a large regional bias, as more west coast schools accepted ACT at this time. Another large source of bias is that schools are more likely to submit their average SAT scores to UNSWR if they are deemed acceptable. Thus the only schools that have submitted all SAT scores to USNWR are those that are on the higher end or meet there historical average. This could result in an overemphasis of the average SAT score since the sample is only taking into account potentially the top SAT average schools. Perhaps the researchers could have added a penalty when a school was missing the data to act as a proxy for potentially withholding worse than average SAT scores rather than dropping them from the study.

The studies mentioned above indicate one over arching theme, which is that the methodology of ranking universities has room for improvement. Although it should be noted that many of the rankings take great care in updating and tinkering with their methodology. Which leads to one of the secondary findings of the studies mentioned above, that the data that many of the rankings rely on might be very noisy, and some of this might be due to the fact that it relies on self reported data by the universities themselves. The study attempts to rectify both of these issues by using objective externally verified sources of information such as Google Scholar metrics, and to incorporate a methodology that might address the noise in the underlying data set.

2.3 Reputational Impact on Rankings and Research Indicators

One of the main objectives of the study is to compare a perception based ranking measured through the USNWR rankings and a research based ranking measured through Google Scholar metrics. Although previous research has indicated that great care should be taken even when using research indicators as a form of quality. Safón and Docampo examined the ARWU and THE World Reputation Rankings in 2019 to examine the impact of reputational bias on research focused rankings.

The purpose of the study conducted by Safón and Docampo in 2020 was to analyze the impact of reputational bias on rankings that use research measures as a substantial part of their methodology [Safón and Docampo, 2020]. Additionally, the researchers were interested in understand if the 'halo effect' (the tendency for an impression created in one area to influence opinion in another area) was present in these rankings and which indicators/criteria it impacted the most.

The study focused on 97 universities from around the globe. Two sets of measures were collected for each of the universities. The 2019 edition of THE World Reputation Rankings was used as the total reputation measure, which consisted of 101 universities that had been determined to be the most reputable through a invitation-only survey. The ARWU Global ranking of Academic Subjects were used as a measure of new reputation. The researchers took a percentile ranking position for each institution within five main areas outlined by AWRU and computed a weighted percentile value based off the publication threshold and the number of institution included in the list and took the average of the five weighted percentile values for each institution to produce the new reputation metric. Additionally the AWRU database was used to find the full time academic staff, which they labeled Staff FTE. Finally, a past reputation measure was obtained through taking the z-score of the natural log of total reputation and subtracting the z-score of the new reputation measure. The calculated past reputation would be the focus of the study moving forward and used to analyze varying metrics within rankings.

The researchers were also concerned with the Matthew effect (the effect of accumulated or compounding advantages [Merton, 1968]) within the nature and science and highly cited measurement. The researchers following previous studies included size in their regression models to control for the size of the department. The two models for nature and science cites and highly cited contained the natural log of the generated past reputation and the natural log of size. The results of these two models showed a significant reputational effect observed in the Nature and Science but did not affect the citations produced by other authors/researchers.

Like many of the other studies mentioned throughout, the sample size is a potential limitation and weakness of the study. Only examining at most a few dozen of the top universities and institutions from each country doesn't exactly paint a clear picture for the median university. As was stated in Meredith's study when he expanded to all 4 tiers and examined the full scope of the university rankings, he found that the 1st quartile or Top 25 behave very differently than the other 75% of universities and colleges. It stands to reason that further self selecting by focusing on only the top 96 global universities further reduces the variation in the data set and these trends may not hold for the other 90% of global universities.

Another limitation observed in the study was the authors ”one size fit all” approach to controlling for size. For example, the authors didn’t think about what counts as size at the university and didn’t define what their metric of size was. There are many universities that rely on part-time faculty, adjunct faculty and teaching professors that contribute much less to research output than a tenure track faculty. The ideal definition of size would be number of full time tenure track faculty. It is entirely possible that some schools rely more on teaching professors or part time faculty and that controlling for size just by looking at the faculty to student ratio wouldn’t fully estimate how many potential research focused faculty the institution contains.

The study showed a key indication in a potential path forward in understanding perception and research output based ranking systems. While external sources such as Nature and Science indicators, or high profile journal publication indicators have evidence of being subjected to reputational bias specifically through the halo effect. Total cites and other raw research based metrics such as h-index and i10-index seem to be more resistant to some of the reputational bias.

2.4 Summary

The research literature proposed thus far indicates three main findings. First, university rankings, specifically the USNWR rankings have a direct impact on admissions decisions and institutional decision makers within the universities themselves. Additionally, the USNWR rankings impact peer perception, whereby current USWNR rankings impact future peer assessment scores, which has the potential to create a feedback loop that is difficult to decouple and favors initially highly ranked universities. Second, the underlying data is generally noisy. But, there are a few key indicators that are important in determining the rank and more emphasis should be put on these indicators that matter rather than focusing on a wide range of criteria that may or may not impact the ranking. Lastly, reputational bias is present even in more objective base research rankings, specifically indicators that rely 'notable' publications. Yet raw citation data seems to be less impacted by reputational or perception bias.

This current study seeks to build off the current literature by examining the difference between a perception based ranking to a research based ranking of graduate statistics departments through the use of raw citation metrics by leveraging Google Scholar total cites, i10-index and h-index. The study will take great care in filtering out faculty members that do not contribute to the graduate statistics department in a research orientated manner to try and accurately measure the size of research producing faculty members. Finally, due to the bias inherent in more established tenure track faculty members citation metrics, the faculty members will be grouped by years of experience and a z-score will be used to standardize each group to effectively measure outstanding impact relative to their peers. A research rank will be established and compared directly to the perception based ranking to measure reputational bias. A bootstrapped confidence interval will be generated and compared to the perception based rankings.

3.0 Data Collection

The data used in the study was collected by using an amalgamation of sources, such as departmental webpages, individual faculty personal web-pages, the mathematics genealogy project, and Google Scholar. First departmental information was collected for each department on the USNWR ranking. Second, unique faculty members descriptive statistics were collected, such as Ph.D. year and Ph.D. granting institution. Lastly, Google Scholar information was collected.

The data was collected from May to June of 2022. The data was subjected to a secondary review in October and November of 2022.

3.1 University Information

A unique id was generated for each university that appeared in the dataset either as a Ph.D. granting institution or as university that has a graduate statistics department in the USNWR ranking. The list contains 1524 universities from across the world, each with a unique 4 digit id. The dataset was constructed by taking the top 250 schools from each region and placing them all in alphabetical order. If a university appeared and it was not on the initial list, it was appended and given the new last digit as its unique id.

3.2 Department Information

The departmental information consisted of name of department, name of all faculty members, number of graduate students and titles of each individual faculty members. The majority of the data was collected via web-scraping using the R package `rvest`. If the information could not be scraped, it was manually added. Additionally, some departments do not provide information on the number of graduate students, for these graduate statistics departments they were given a value of 0.

At the end of this process each row was an individual faculty member with a title, and a unique id based of the university of their graduate statistics department. Additionally, each department had a number of graduate students associated with their unique id.

3.3 Faculty Information

The individual faculty information consisted of Ph.D. year, Ph.D. granting institutions, Ph.D. field/discipline, year of first tenure track position. This information was collected in multiple stages and passes using different sources. The first pass was to scrape the information if possible from the departmental web-page. If the departmental web-page had a high level of abstraction and scraping was not a viable solution, manual retrieval of the information was done. The vast majority of the faculty level information was collected manually.

But there was some situations in which either the departmental web-pages lacked sufficient information or had no information at all. When this occurred we used a multitude of methods to retrieve the data. The first effort was made to identify a personal web-site for individual faculty members. If this method did not provide the necessary information, the mathematics genealogy project database was used to search for individual faculty members. If after these steps no information was retrieved for a faculty member, an NA was placed in each of the missing or relevant variables.

4.0 Methods

The USNWR rankings are the preeminent university ranking used within higher education within the United States. The rankings have both internal organizational reputation within universities and external influence on application and admission decisions to students.

Research conducted on the rankings previously have shown most movement year to year is mostly due to noise and that universities stay relatively stagnate indicating the potential for bias. Specifically, the USNWR graduate department rankings rely solely on self-reported survey data. This type of data is prone to bias in the form of selection bias, order-based bias, and social desirability bias.

The study relied on collecting data for each professor to construct a representative sample for each department to measure each universities research output to compare to the USNWR statistics department ranking. The research output for each professor was measured with the three main Google Scholar research metrics, total cites, i10-index, and h-index. In addition PhD granting institution, years of experience, PhD degree type and title were collected to try and examine the types of bias present in the rankings. The data was cleaned and standardized in multiple methods and a bootstrap sample was taken of the cleaned and standardized datasets. Finally, an error metric or an MSE like statistic was calculated against the USNWR statistic graduate department rankings.

4.1 Participants

The researcher relied on a convenience sample based of the existing universities on the USNWR statistic graduate department rankings. The USNWR constructed their ranking by using a list of all statistics doctoral granting universities in the United States provided by the American Statistical Association (ASA). Out of all the doctoral granting universities provided to USNWR by the ASA, 101 were included in the ranking. The USNWR did not include any university that received less than 10 rankings from their peers.

The individual participants included 1003 tenure track statistics professors at one of the 101 universities included in the USNWR ranking. The individuals were selected because they had a Google Scholar profile and meet the criteria for being counted as a statistics tenure faculty member.

4.2 Procedure

The procedure for this study relied on two main components, data collection and data analysis. But after the data collection and before the data analysis could take place a criteria for determining which faculty members to include in the study needed to be determined by the researcher.

The procedure for determining what faculty members to include in the study was broken down into three main components or steps. First, determining whether the university on the USNWR ranking had a stand-alone statistics department, or an interdisciplinary department. Second, filtering non-tenure track faculty that are not apart of the statistics department or the group of statistics focused tenure faculty in an interdisciplinary department. Third, determining whether each individual tenure track faculty had a Google Scholar profile.

Each university was determined to either have a stand-alone statistics department or an interdisciplinary department based off the name of the department, the titles of the tenured faculty members, and the presence of a unique statistics departmental web page. Therefore, departments were classified as being in a stand alone statistics department or an interdisciplinary department and all individual faculty members were similarly labeled based off their university.

For example, the University of Pittsburgh has a titled Statistics Department with a specific departmental web page and the tenure faculty are Department of Statistics professors. Thus, the University of Pittsburgh was classified as having a stand-alone statistics department. The University of Notre Dame's Ph.D. in Statistics is granted by the Applied and Computational Mathematics and Statistics Department. The Applied and Computational Mathematics and Statistics Department contains the Ph.D. in statistics, thus, the department was classified as an interdisciplinary department because of the title and the presence of non-statistics tenure faculty within the same department.

Once the labels for each graduate statistics department were made, the next step in the process was to screen and filter the individual faculty members. If universities contained a stand-alone statistics department, then a professor must be a tenured track faculty member and their primary appointment must be a part of said statistics department. For example, visiting professors, adjunct or teaching faculty were not included in the ranking.

If the university did not contain a stand alone statistics department or it wasn't clearly distinguishable from the department or school website, then the following process was followed. If the professor had a Ph.D. in statistics and was a tenure-track faculty member, then they were included. But if the faculty member were a part of an interdisciplinary department and did not have a Ph.D. in statistics then they were not included in the study.

The next step in the professor selection process was filtering out professors if we could not ascertain the relevant and necessary information to be able to conduct the analysis. Every professor needed to have a Google Scholar profile/Google Scholar id. Google Scholar id's were collected using SerpAPI's Google Scholar API and then appended to each faculty member.

Once the Google Scholar id's had been appended to all of the faculty members in the dataset, the R package function scholar was used to collect the total cities, i10-index and h-index for each faculty member.

4.3 Data Pre-Processing

The faculty members that received a Google Scholar id were grouped into 30 bins based of the number of years of experience they have from their first tenure-track faculty position. The experience bins were determined based off the frequency distribution of faculty members, such that each bin had roughly the same amount of faculty members within them while not splitting within a year which was not possible given the data the researcher collected. The smallest bin had 21 faculty members while the largest bin had 54 faculty members.

The z-scores/normalization scores for each of the 30 bins were calculated for each of the three Google Scholar research metrics, total cites, i10-index and h-index. Each bin had their group mean and standard deviation calculated for total cites, i10-index and h-index. Then each faculty member within the group had their three research metrics normalized based their within group label.

$$z_{j,i} = \frac{x_{j,i} - \overline{x}_j}{\sigma_j} \tag{1}$$

Table 1: Bin number and years of experience and each respective bin mean and standard deviation used to calculate each faculty members normalized z-score for all three Google Scholar metrics.

Google Scholar Normalization Scores			
Bin Number (Years of Experience)	total cites (mean, sd)	i10 index (mean, sd)	h-index (mean, sd)
1 (56 to 46)	55645.14, 93772.98	180.33, 178.81	69.29, 55.03
2 (45 to 42)	21791.13, 28215.49	107.19, 74.91	46.97, 27.72
3 (41 to 39)	11582.86, 11579.37	86.91, 69.31	40.09, 22.50
4 (38 to 37)	55938.52, 117015.45	138.05, 169.52	55.24, 50.75
5 (36 to 35)	18729.76, 20597.45	98.36, 64.87	46.12, 25.03
6 (34 to 33)	9591.44, 10877.72	73.41, 55.43	36.30, 20.19
7 (32 to 30)	18322.36, 29784.84	96.30, 70.14	43.79, 24.08
8 (29 to 28)	11010.57, 12310.67	93.71, 85.54	39.93, 19.99
9 (27 to 26)	7720.14, 7528.22	76.83, 60.74	35.97, 19.42
10 (25 to 24)	20221.92, 38637.52	82.08, 80.57	39.96, 32.75
11 (23 to 22)	9377.26, 12436.11	67.11, 50.73	34.26, 18.46
12 (21 to 20)	7553.97, 12029.62	57.49, 52.07	29.72, 17.50
13 (19)	5842.25, 6084.40	58.00, 36.40	32.13, 14.32
14 (18)	6671.00, 19146.25	44.79, 39.55	25.94, 17.03
15 (17)	6179.23, 13736.91	43.44, 27.11	24.49, 10.09
16 (16)	4454.56, 6632.72	46.22, 44.96	24.06, 11.87
17 (15)	4313.50, 6848.39	41.15, 23.42	24.08, 9.40
18 (14)	2476.10, 3697.77	25.79, 15.44	17.97, 8.42
19 (13)	2464.06, 3569.58	29.97, 25.81	18.31, 10.68

20 (12)	3390.63, 7037.20	26.67, 17.10	19.54, 10.04
21 (11)	1832.87, 2012.49	24.72, 18.15	17.15, 8.58
22 (10)	2838.23, 8632.96	24.67, 18.86	17.47, 9.89
23 (9)	1756.26, 3076.41	18.70, 15.15	14.26, 7.66
24 (8)	1331.77, 3873.38	15.64, 16.02	12.50, 8.55
25 (7)	1166.55, 1320.84	17.43, 14.61	13.60, 7.46
26 (6)	650.89, 644.13	12.05, 8.39	10.66, 4.66
27 (5)	643.59, 1383.83	9.74, 10.74	9.11, 6.56
28 (4)	640.10, 1174.52	9.44, 9.57	8.96, 6.00
29 (3)	436.23, 714.84	7.85, 7.10	7.73, 4.54
30 (2 to 0)	293.29, 415.02	5.41, 4.62	6.29, 3.27

4.4 Research Rank Statistic

The researcher was primarily interested in determining if there was any deviance between perception of graduate statistic departments measured through the USNWR rankings and a research output based ranking for programs within the United States. In order to conduct this comparison, the researcher created a metric or statistic to measure the research output of each graduate statistics program. The metric, which used the standardized Google Scholar metric scores based off years of experience, was used as the basis to generate a statistic called research rank, which would be used as a direct comparison to the USNWR peer assessment survey score. The USNWR ranking acts as a proxy for perceived quality of a graduate statistics program compared to its research based output.

The process of creating the research rank statistic utilized the three Google Scholar metrics for each department. The median standardized scores for each of the three Google Scholar metrics, total cites, i10-index, and h-index were taken for each department for the relevant faculty members that belong to each respective department. For each standardized Google Scholar metric each department was ranked by the value of the median faculty value for each individual metric. This resulted in three rankings based off the median value for each department. The research rank was then determined by taking the mean of the three rankings and then ranking the departments from lowest to highest.

4.5 Error Measurement

We measured the MSE for each of the three Google Scholar metric ranks, and used the USNWR ranking as the true error measurement. Thus the MSE was a measure of the difference between the research output based ranking for each Google Scholar metric and perception.

$$MSE_{i,r} = R_{i,j} - R_{i,k} \tag{2}$$

where i = department {1,2,...,97}, j = research output ranking, k = USWNR ranking, r = Google Scholar metric {1,2,3}

After the MSE was calculated for each department for each Google Scholar metric, the RMSE was calculated for each department.

$$RMSE_i = \sqrt{\frac{\sum(MSE_{i,r})^2}{3}} \tag{3}$$

where i = department {1,2,...,97}, r = Goolge Scholar metirc {1,2,3}

The last measurement of error was perception bias statistic. This was the difference between research rank statistics and USNWR ranking. This statistic was both positive and negative. If a value was negative it indicated a high perception bias and if it had a positive value it meant their research output ranking exceeded their perception ranking.

4.6 Bootstrapping

We drew 1000 bootstrap samples of each department to generate an estimate for the range of values each department could take. The process for generating each bootstrap sample was:

1. Sample with replacement the faculty from each university. This meant finding the unique number of faculty in each department and drawing a sample of that many faculty members from the data set for each respective department
2. Calculate the median value for each of the three Google Scholar metrics for each sample drawn for each department
3. Calculate the new ranks based off the bootstrap Google Scholar metrics for all three Google Scholar metrics
4. Calculate bootstrapped research rank based off the 1000 Google Scholar metric median values.

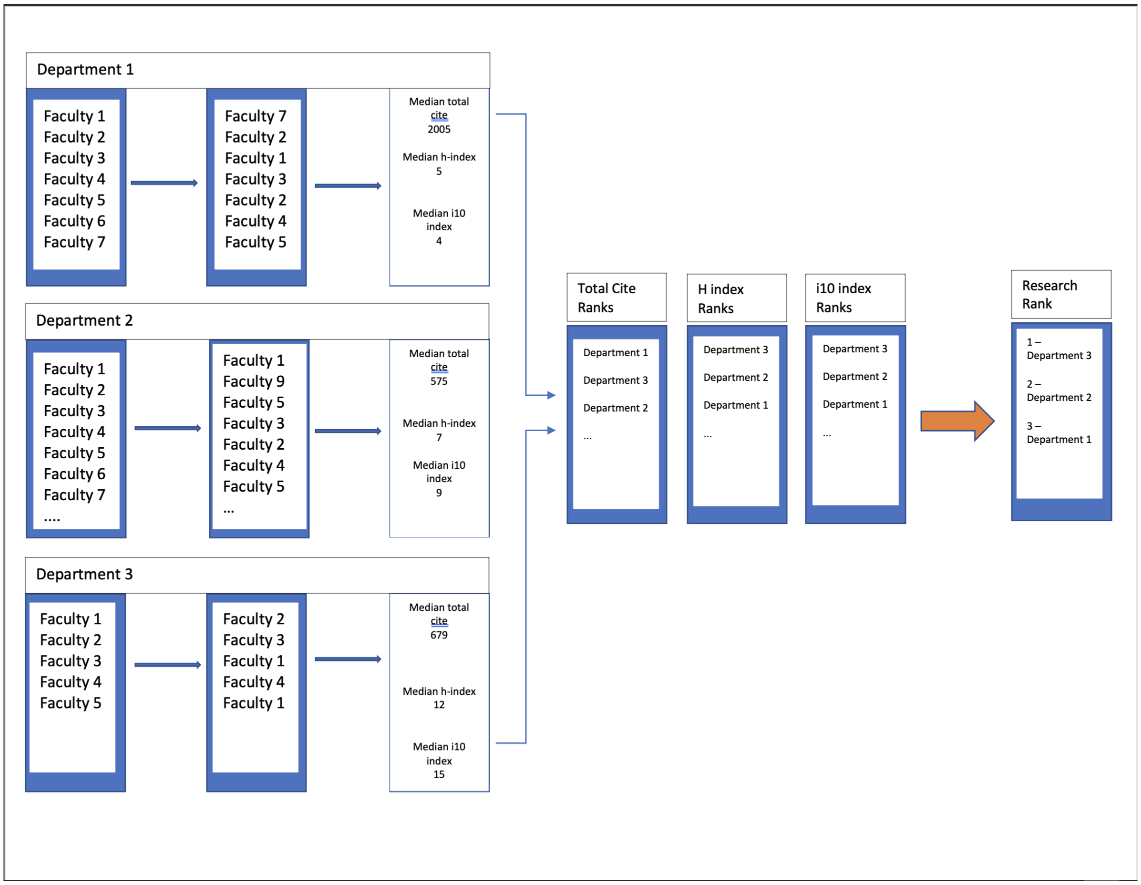

Figure 1: Bootstrap sampling of department and faculty to generate research rank

5.0 Results

The main research question of the study was to examine the difference between perception of statistic departments measured through the USNWR peer assessment survey data ranking, and a research output based ranking measured through Google Scholar metrics. Additionally, the researcher examined multiple metrics based off the research rank statistic and compared them to the perception proxy in the USNWR rankings. 1000 bootstrap samples were collected to estimate the ranges of the research ranks for each university and then compared to the perception based rankings.

5.1 Perception Ranking and Research Ranking

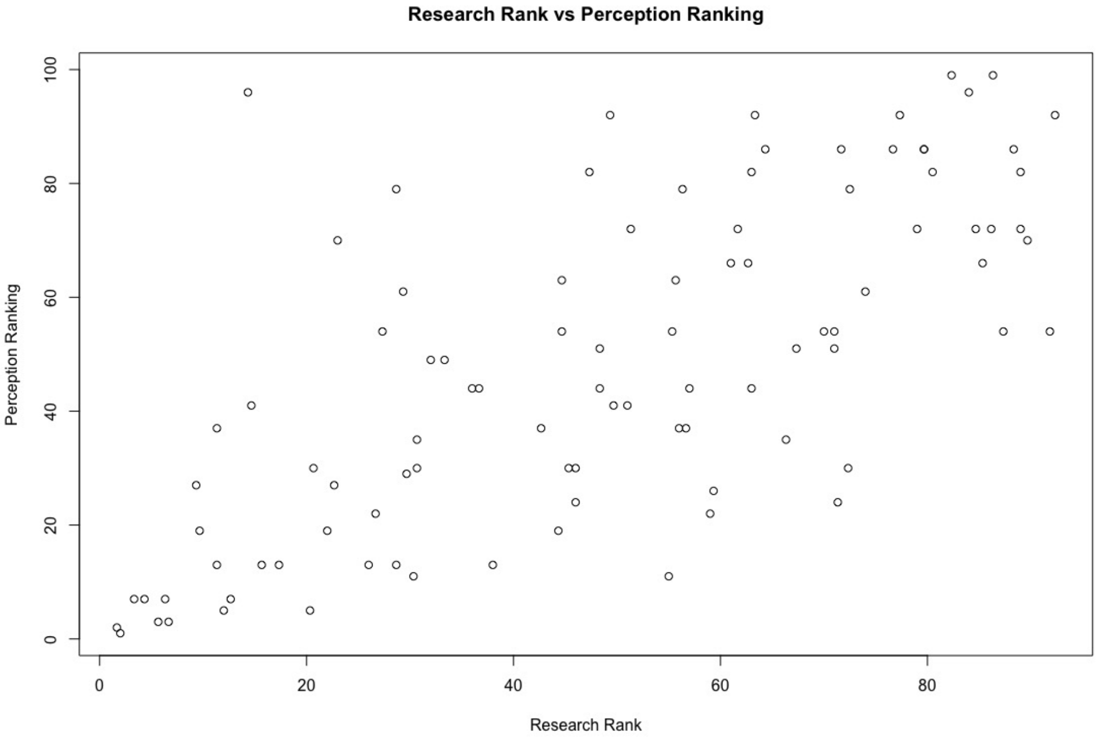

Figure 2: Research Rank vs Perception Ranking

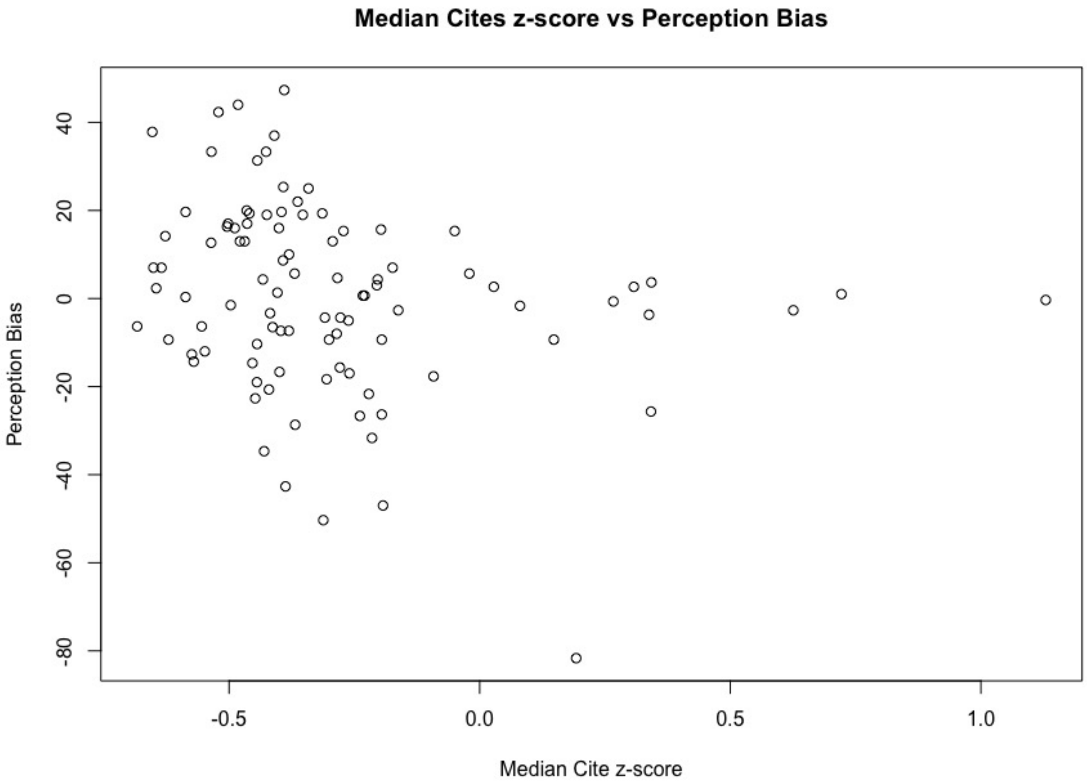

Figure 3: Median z-score total cites by department plotted against perception bias

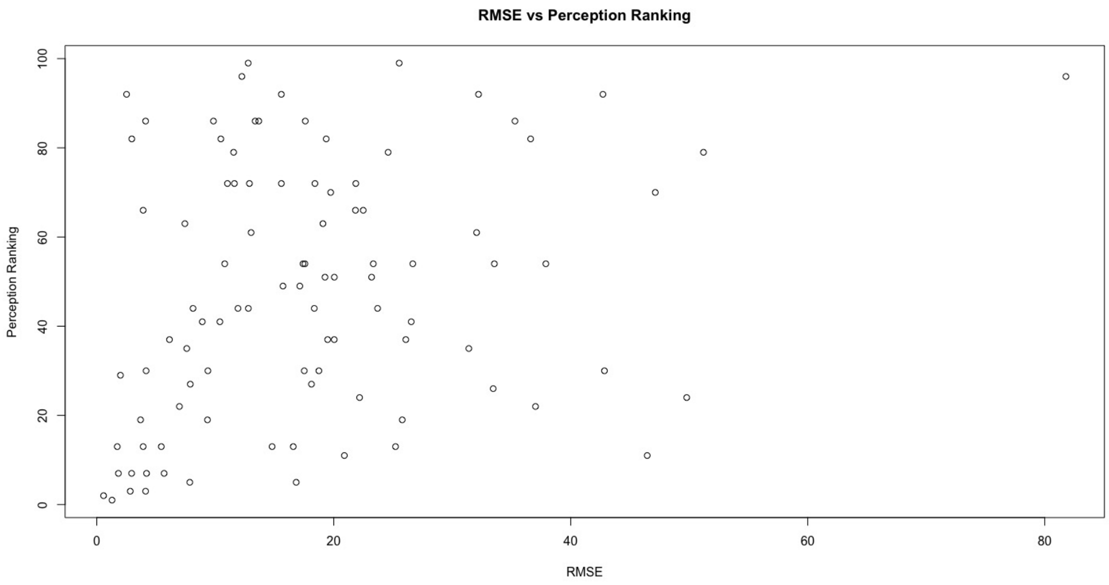

Figure 4: RMSE vs Perception Ranking

5.2 Total Cite z-score and Ranks

The total cite z-score was calculated using Equation 1 with the values from Table 1. The correlation between highest cite z-score per department and perception rank per department was -0.66. The correlation between highest cite z-score per department and research rank per department was -0.65. The correlation between median cite z-score and perception rank was -0.58. The correlation between median cite z-score and research rank was -0.81.

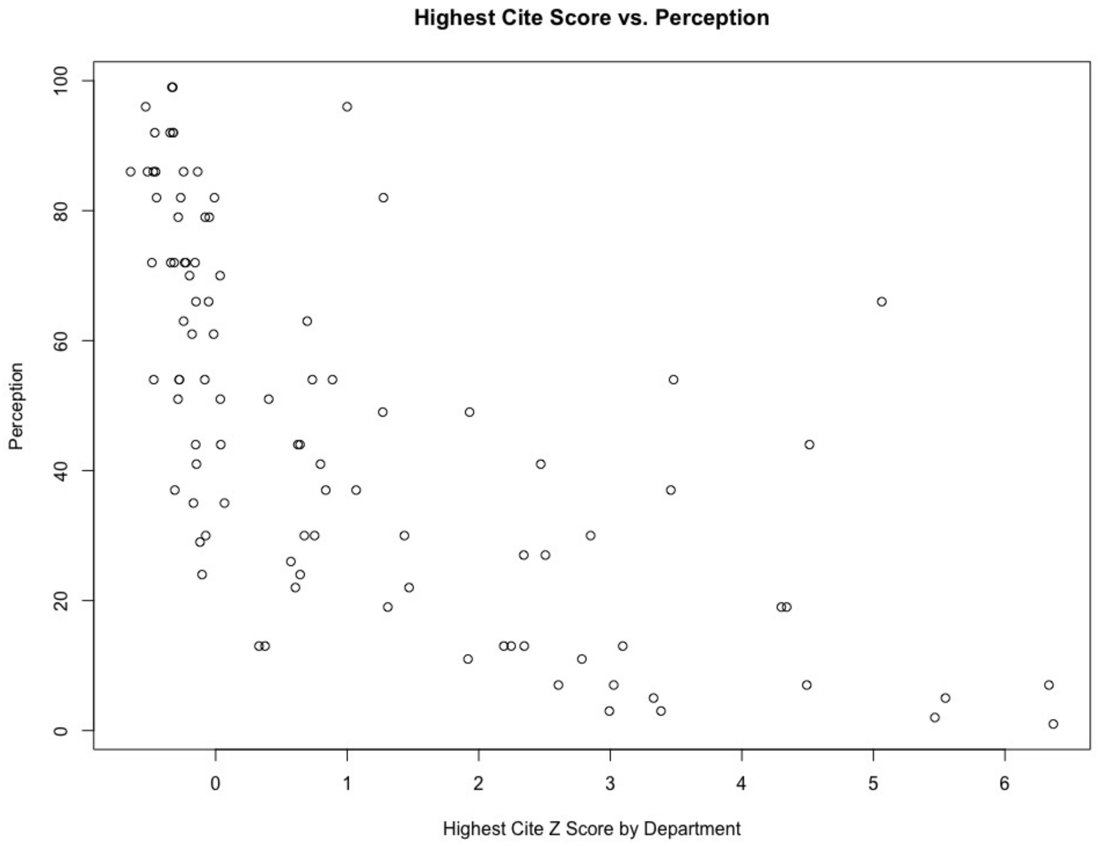

Figure 5: Highest Cite z-score of each department plotted against perception ranking

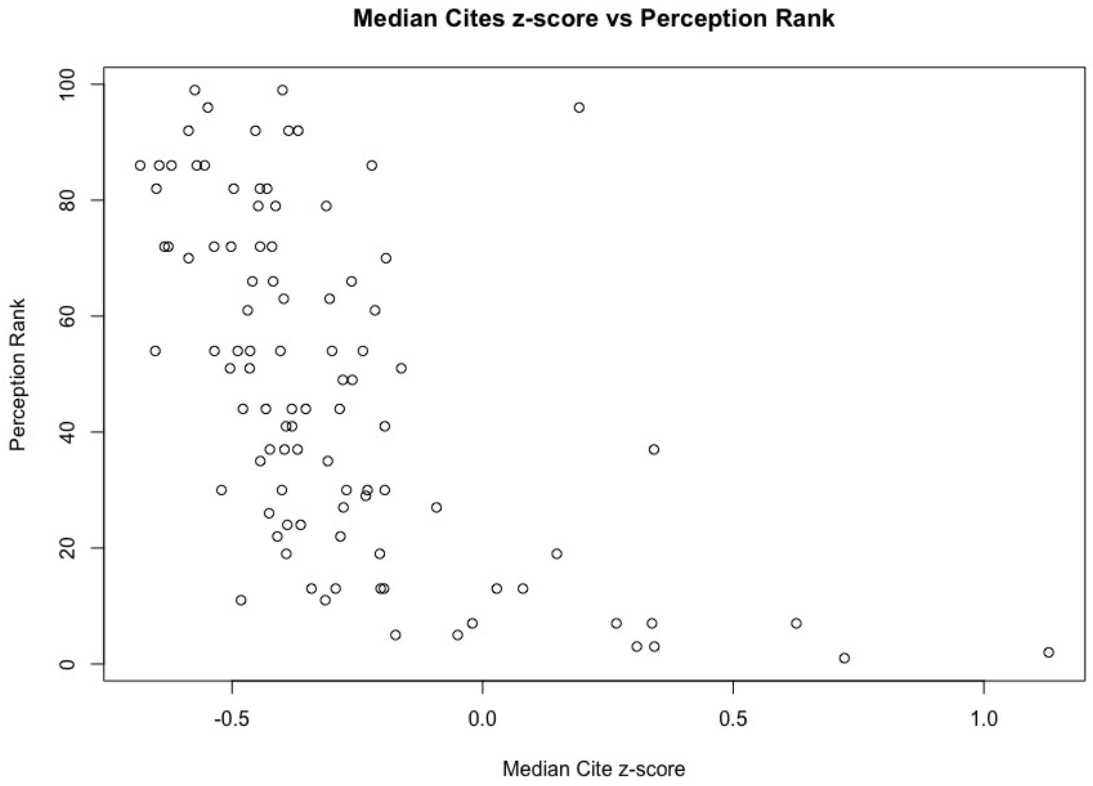

Figure 6: Median Cite z-score of each department plotted against perception ranking

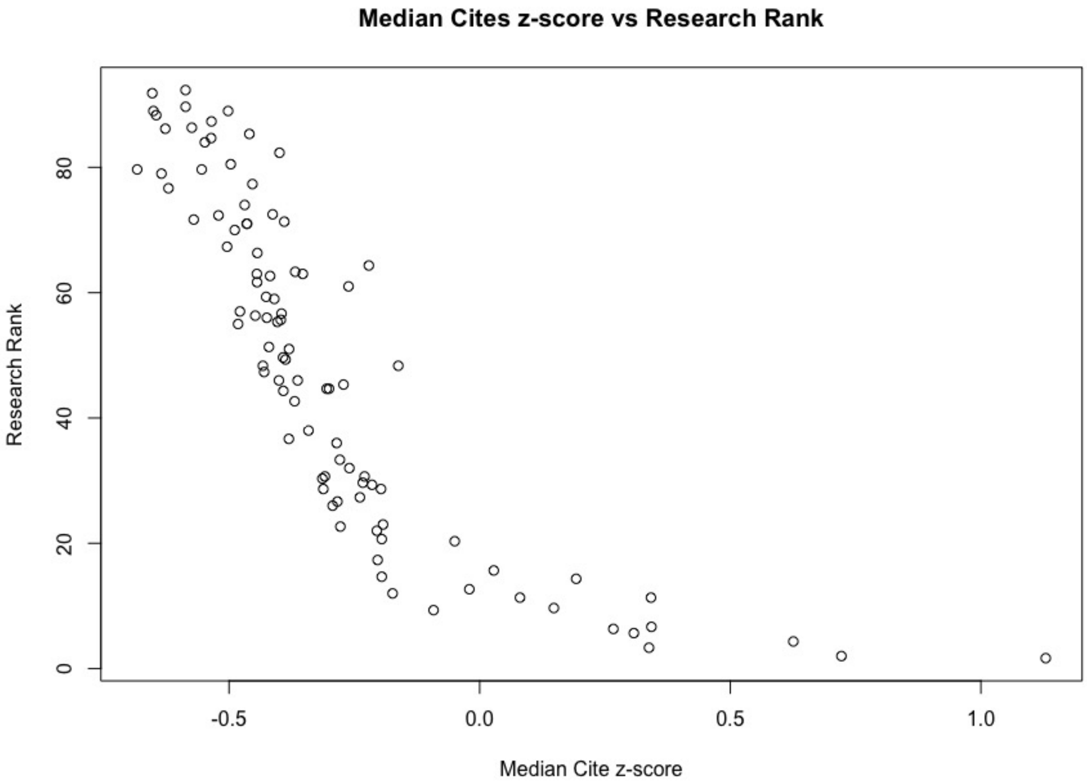

Figure 7: Median Cite z-score of each department plotted against research ranking

5.2.1 H-index z-score and Ranks

The h-index z-score was calculated using Equation 1 with the values from Table 1. The correlation between highest h-index z-score per department and perception rank per department was -0.76. The correlation between highest cite z-score per department and research rank per department was -0.73. The correlation between median h-index z-score and perception rank was -0.70. The correlation between median h-index z-score and research rank was -0.92.

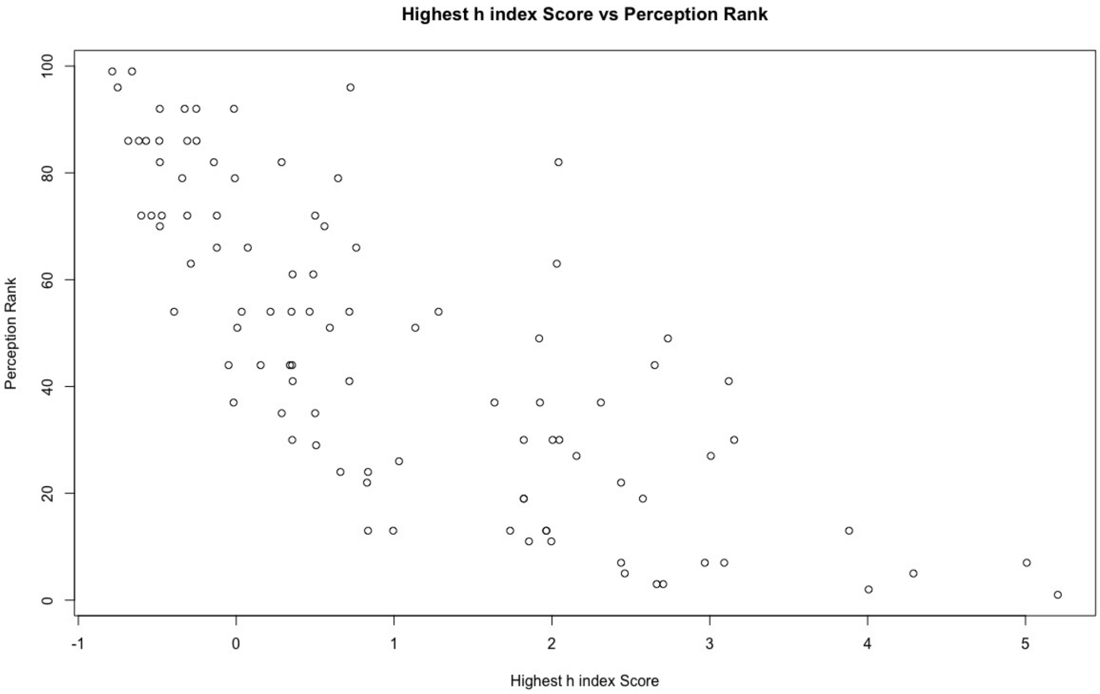

Figure 8: Highest h-index z-score of each department plotted against perception bias score

5.3 i10 Index z-score and Ranks

The i10 index z-score was calculated using equation 1 with the values from Table 1. The correlation between highest i10 index z-score per department and perception rank per department was -0.77. The correlation between highest cite z-score per department and research rank per department was -0.74. The correlation between median i10 index z-score and perception rank was -0.67. The correlation between median i10 index z-score and research rank was -0.91.

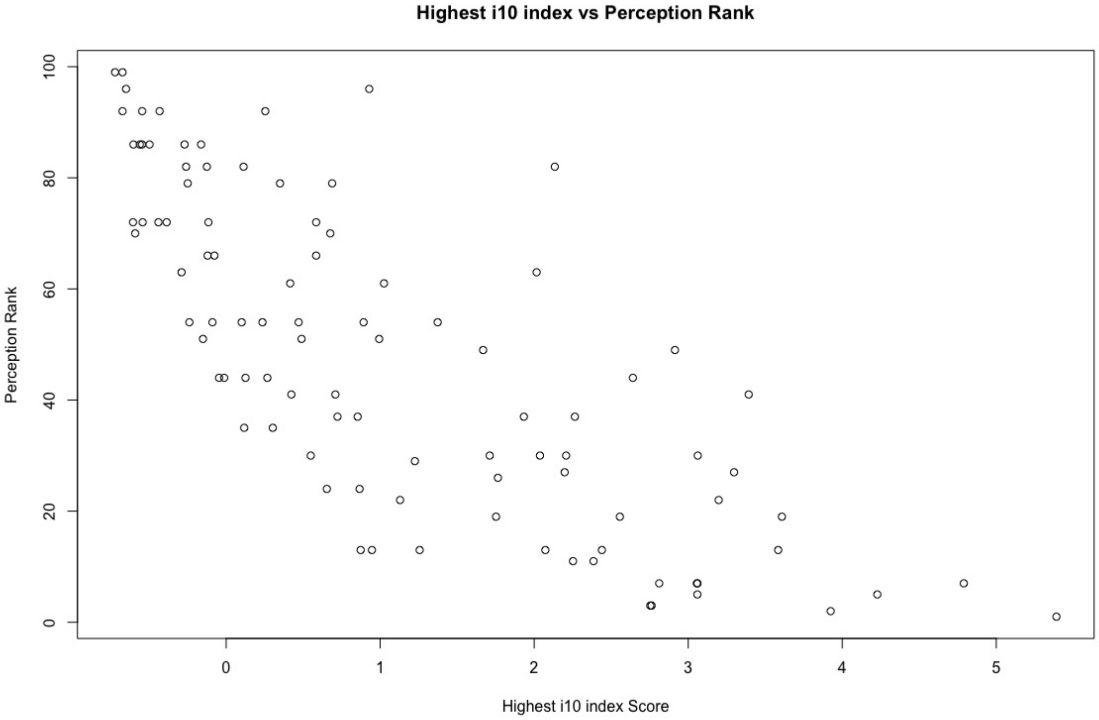

Figure 9: Highest i10 index z-score of each department plotted against perception bias score

5.4 Number of Professors and Rankings

The correlation between number of professors and perception rank per department was -0.79 and -0.52 for research rank.

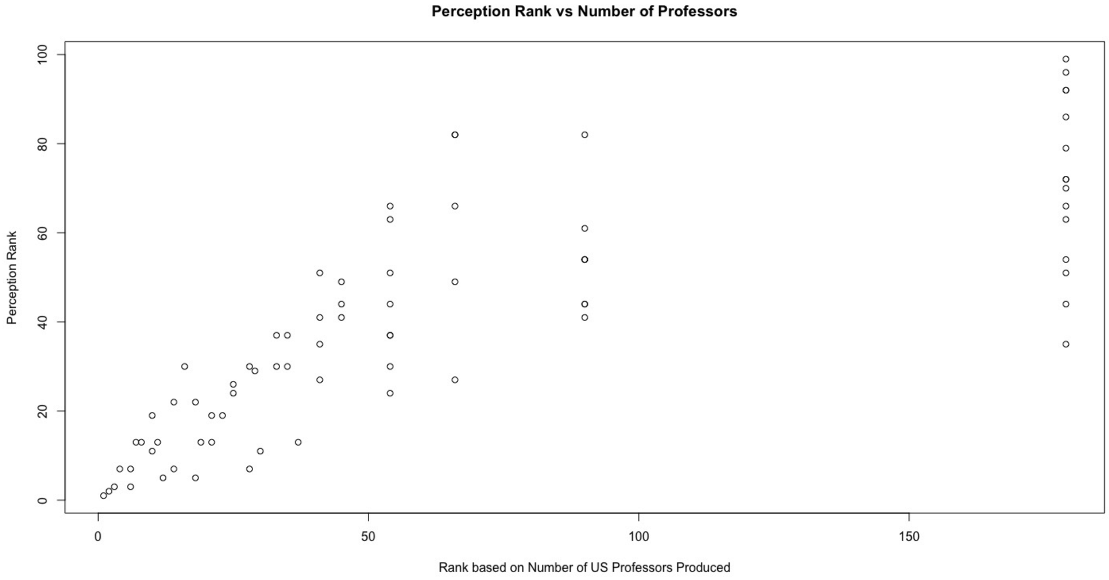

Figure 10: Number of Professors plotted against perception rank

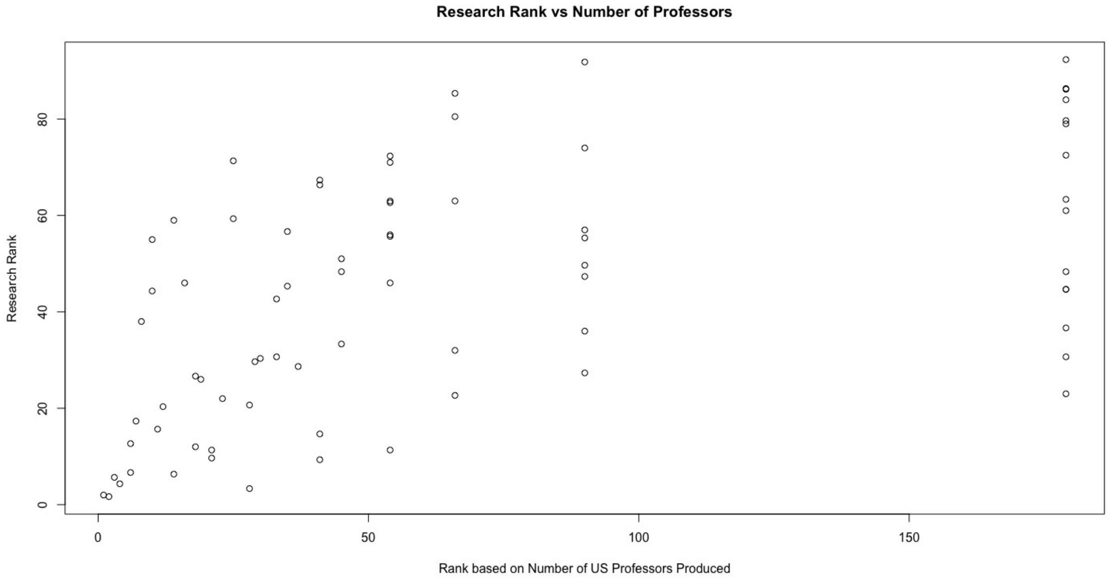

Figure 11: Number of Professors plotted against research rank

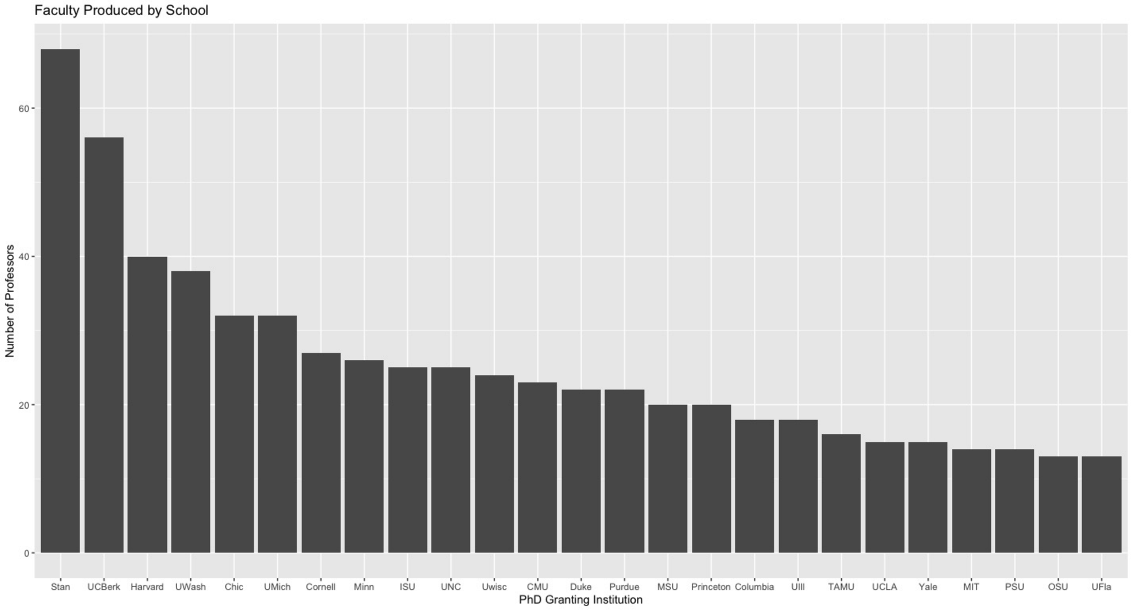

Figure 12: Number of Faculty Produced by Department

5.5 Bootstrap Metrics

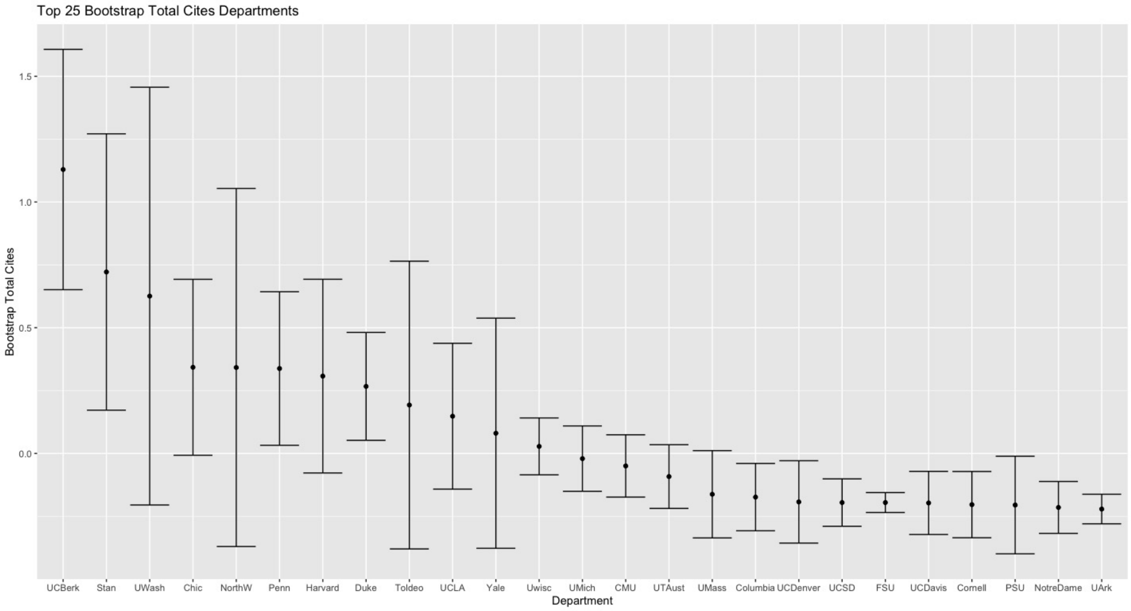

Figure 13: The median of the total cites z-score across 1000 bootstrap samples. The error bars are the standard deviation of the total cites z-score for the bootstrap sample. The departments are ordered from 1st rank to 25th rank of median total cites z-scores.

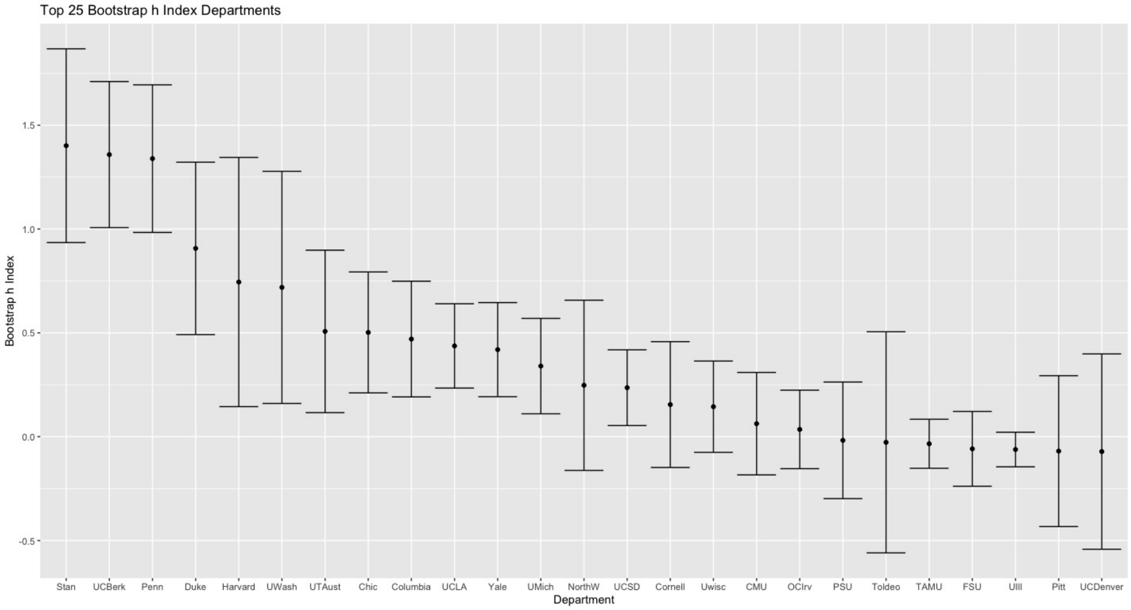

Figure 14: The median of the h-index z-score across 1000 bootstrap samples. The error bars are the standard deviation of the h-index z-score for the bootstrap sample. The departments are ordered from 1st rank to 25th rank of median h-index z-scores.

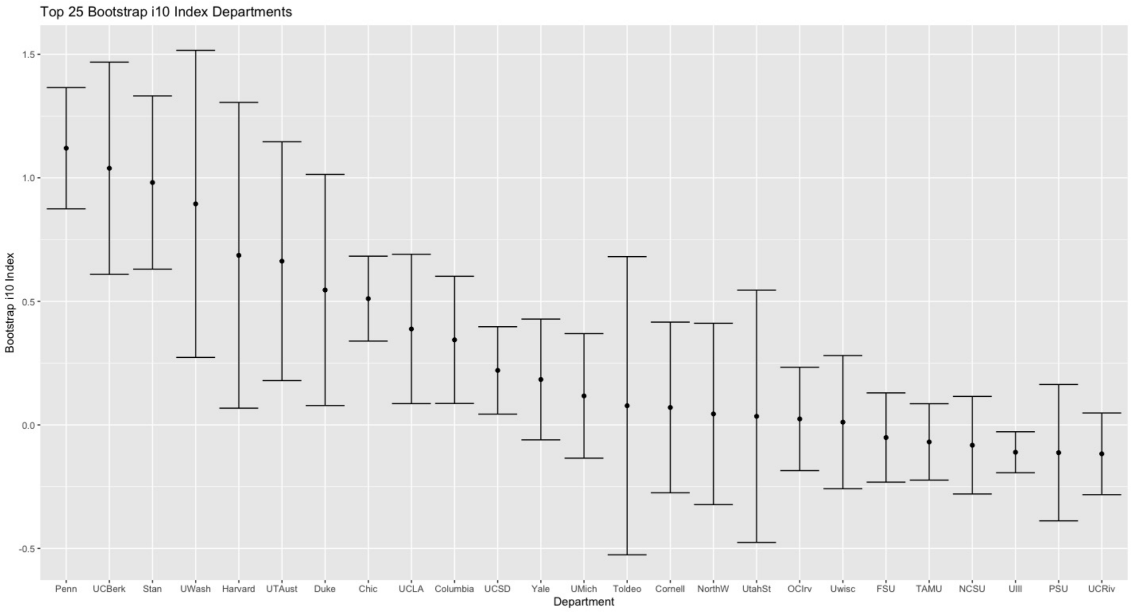

Figure 15: The median of the i10-index z-score across 1000 bootstrap samples. The error bars are the standard deviation of the i10-index z-score for the bootstrap sample. The departments are ordered from 1st rank to 25th rank of median i10 index z-scores.

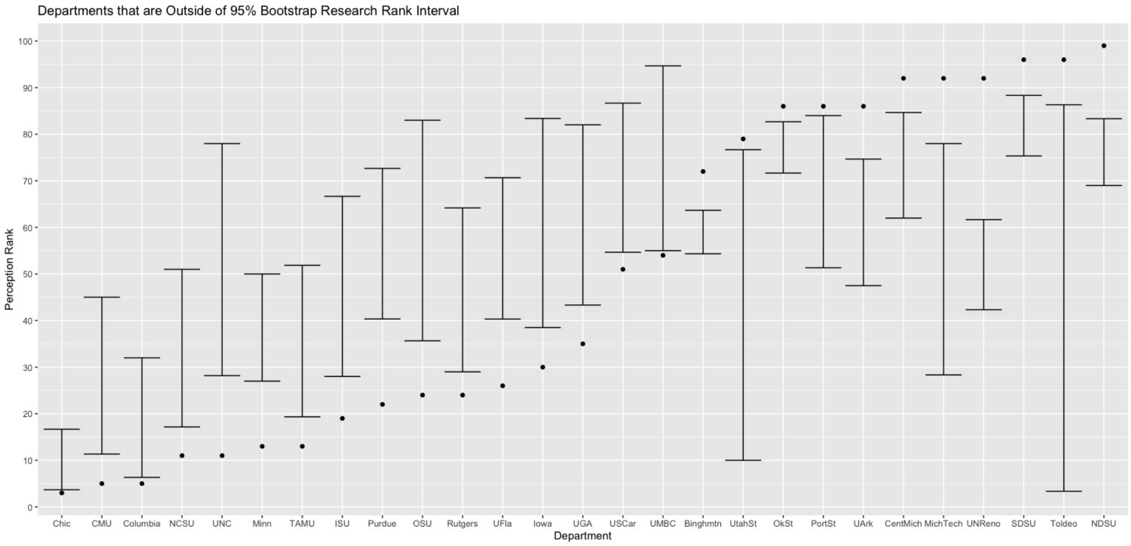

Figure 16: 1000 Bootstrap 95% confidence intervals for Research Rank. Schools included in the chart have a perception ranking that falls outside of the confidence interval. The points are the perception ranking and are added for context. The departments are ordered from highest perception rank to lowest perception rank.

6.0 Conclusion

6.1 Summary of Findings

University rankings have become an important part of the admissions processes and their audience has grown from curious faculty members, then to students, universities and governments. Most of these rankings target undergraduate institutions and use about a dozen ranking criteria. Yet there is still a lack of coverage for graduate programs outside of business, law and medical school. One of the eminent rankings is the USNWR, which has begun ranking graduate programs through peer assessment survey data which relies heavily on institutional reputation. With the main focus of graduate students publishing and collaborating on research, it seems reasonable to build a ranking centered around department research output.

Data was collected for every tenure statistics focused professor at the 96 graduate statistics departments. The data was normalized by years of tenure to help mitigate the inherent bias in the Google Scholar metrics. A research rank statistic was generated for each department based off the mean of the normalized Google Scholar metrics for each professor within their department. Lastly, bootstrapping was conducted and was used to calculate the plausible ranges of values departments could take. This was used to determine if a department deviates measurably between its research rank and perception based rank.

The results suggest four main findings:

1. There is a measurable difference between the perception based ranking and the research output based ranking.
2. Perception based ranking may rely on the highest performers.
3. Larger departments seem to benefit from a perception based ranking system, while smaller departments benefit from the research based output ranking.
4. There is a small group of elite graduate statistics departments in terms of research output, with the rest of the departments showing surprising parity. This indicates a tiered approach may be more beneficial and informative than a list approach.

6.2 Conclusion

Based off Figure 2 perception ranking and research ranking have a moderate relationship. This is roughly what we expected because the top schools are likely to be accurately ranked. But as the ranking decreases the two rankings were more likely to be divergent. This can be seen in Figure 2, where the first dozen or so universities have little if any perception bias. But outside of these groups of schools there tremendous deviance between the median cite z-score and the perception. This same general trend plays out in Figure 3, where the bottom left corner shows a group of schools more accurately ranked between both rankings. But outside of this very small bottom left corner, there is very little consistency.

One of the recurring themes throughout the results was there are clearly a small group of highly ranked graduate statistics departments both in terms of research and perception. But, outside of this initial group there is surprising amount of parity especially when considering the median research metrics for each department.

For example, the median h-index z-score for University of California, San Diego is 0.24, which ranks 14 and Cornell has a median h-index z-score of 0.15, and ranks 15 in median h-index z-score. Yet, University of Washington has a median h-index z-score of 0.72 and ranks 6, while the 7 ranked University of Texas, Austin has a median h-index z-score of 0.51. Based off the rankings, University of Texas, Austin would assumed to be closer to University of Washington than University of California, San Diego, yet the true scores indicate it is as far from the 6 ranked university as it is to the 14 ranked university.

In contrast examining the highest total cite z-score for each department the correlation between the perception ranking is -0.66, the negative sign is due to using a ranking so the lower the score the better the ranking. Figure 5 shows a fairly moderate trend between the highest total cite z-score and the perception ranking. This trend plays out across all three Google Scholar metrics. Examining Figure 8 h-index and perception rank and Figure 9 i10-index and perception rank shows an even stronger correlation. These seem to indicate that perception is strongly based off the assessment of the top performer within each department.

When compared to the median total cite z-score, the correlation is reduced from -0.66 to -0.58 for the perception based ranking. The research based ranking increased from -0.65

to -0.81 as can be seen in Figure 6 and 7.

The other two metrics had very similar trends, there is a decrease in correlation for the perception based ranking when moving from highest z-score metric to median. Meanwhile the research based ranking would move in the opposite direction having a stronger correlation for the median z-score metrics than the highest z-score metrics.x

Another key factor that contributes to the difference between the two ranking systems used in the study is that large departments seem to fair better in a perception based ranking rather than a research based ranking. Figure 10 shows a strong correlation between perception rank and size of department measured in tenure track faculty, where the correlation is -0.79. In Figure 11, there is a lack of emphasis on size of the department, where the correlation is -0.52.

An intriguing chart produced from the dataset is Figure 12, where it shows the number of faculty produced by school. As the dataset is open source and will be made available, this information could be used in future studies to examine quality of graduate statistics departments by measuring how many faculty the graduate program produces. Additionally, future research could examine the ratio of doctoral degrees granted to faculty produced. It is also important information for graduate students focused on becoming a tenure tracked faculty member.

The last aspect of the study leveraged bootstrapping the original dataset 1000 times with replacement. The reasoning behind this was to try and estimate the range of values a university could take and compare those range of values to the perception and research based ranking. Figures 13, 14 and 15 show roughly the same overall trend, there is a clear group of schools that even when taking their median value and subtracting one standard deviation would still rank higher than the other 10 or 15 schools in the Top 25.

When looking at Figure 14 and Figure 15 in particular, the top 3 universities are the same, but in different orders. Additionally, when examining h-index the top 3 have a higher lower bound than the 4th department and in regards to i10-index the top 3 have a higher lower bound than the 5th ranked school.

One interesting aspect of Figure 13 is that when you examine the first 11 departments, the bounds of one standard deviation are much larger than the remaining 14 departments.

The researchers hypothesis is that this is due to the few outliers/high performers/high researchers within these departments. These very few faculty members produce significantly more research and highly cited researcher than the majority of faculty members, even within their own department and are the reason the scores are so high for their departments. This is a situation where the perception ranking does very well, and is why there is very little difference between the perception top 10 or so and the research ranking top 10 or so. The two rankings through different metrics are essentially measuring the same thing. It is when the rankings move outside of this top 10ish universities that there is significant divergence.

Lastly, the research did a 1000 bootstrapped sample with replacement and then recalculated each category of Google Scholar metric which enabled the researcher to generate a bootstraped 95% percentile interval for the research rank. The full results are shown in Appendix E, but Figure 16 is a chart of only the departments that have a perception ranking outside of the 95% bootstrap percentile interval for their research rank. When Figure 16 is taken into consideration with Appendix E, then there seems to be a pattern. Out of the 16 departments that have a higher perception ranking than the lower bound of their 95% percentile interval only 5 are not in the top 25 largest departments and all but 1 fall in the top 50 largest departments.

Figure 16 demonstrates two types of situations. Departments that seem to be reasonably far away from their 95% credible interval and have a reasonably narrow bound, indicating they may be significantly miss ranked based off perception. This would include departments such as Binghamton and University of Nevada, Reno. Two, departments seem to be close to their 95% credible interval, but have a large band, this would include departments like Utah State and University of Maryland, Baltimore County, which seem to indicate perhaps somewhat accurate perception based ranking due to the large variance.

6.3 Limitations

There were three main limitations identified during the planning phase of the study conducted. First, we only used the same schools ranked by USNWR. Second, Google Scholar was selected as the only research metric database to use. Thus any professor without a Google Scholar profile was dropped from the study. Third, the data about professor titles and which departments they belong to relied on the university webpages themselves which have varying levels of accuracy outside of the control of the researcher.

We decided to use the same schools ranked by the USNWR because the main scope of the study was to measure the deviance between a research based output statistic graduate department ranking and perception of a statistics department measured through the USNWR rankings. Yet, this is a form of selection bias in and of itself. Ideally, with more time the researcher would collect every R1 institution and design a unique survey to quantify perception and compare them all within the same dataset.

There are numerous ways to measure research output by an individual professor. Google Scholar was selected to focus on during the study because it offered ease of access, the largest amount of professors participating within its ecosystem, and provides the total cites, i10-index and h-index for every profile on their platform. Additionally, due to time constraints it was not feasible to collect every platforms research metrics. A problem also arose when trying to aggregate different research metrics from different platforms. For example, aggregating into one dataset the ResearchGate RG/RI score and the three Google Scholar metrics would have resulted in loss of information and was unwieldy.

Only using Google Scholar was the largest limitation in the study because it introduced selection bias. The people most likely to use Google Scholar are the professors most likely to care deeply about research and publishing, thus they will tend to have higher Google Scholar metrics. For example, some departments had low Google Scholar usage, resulting in only a few professors representing entire departments. Meanwhile, four departments had no professors using Google Scholar. In the study these departments were dropped and is why the final study examined only 97 of the 101 universities that USNWR ranked.

The information specific to professors such as PhD granting institution, PhD year, start year and title was all collected off personal websites and departmental websites manually. This information is self reported and is very difficult to validate. Occasionally, the researcher did notice conflicting information and the best educated guess was made in trying to rectify conflicting personal website vs departmental website information. But these occurrences were rare overall.

Lastly, we decided to rank each graduate statistics department by the research output. This does not mean that research output is synonymous to quality of department. The research based ranking done throughout this study is not meant to convey quality or lack thereof. There are many aspects and ways to determine quality in a graduate education. Research output was selected by the researcher for this study because it offers an objective and comparable way to measure departments. Additionally, the main of objective of many graduate students is to conduct research and publish papers. While this is certainly not the only reason to pursue a graduate education it is perhaps one of the largest factors.

During the course of the study some potential limitations arose that were not anticipated during the design aspect of the study. First, 11 departments had 3 or less faculty members and 31 departments had 5 or less faculty members. These small sample sizes are less than ideal, as it magnifies and gives these faculty members more weight than in a department that is larger. Thus, if this professor was an outlier, then it could skew that specific department heavily. An example of this is Toledo University, one of only two professors is the 3rd most total cited professor within the dataset. This is surely not an accurate picture of the entire department as it is highly unlikely that the other professors would be as highly cited.

Another major limitation was generating the bootstrap samples. There has been some recent research indicating that when constructing an empirical ranking using a bootstrap sample the asymptotic principals for an n-out-of-n bootstrap do not hold [Hall and Miller, 2009]. Additionally, that the confidence interval constructed from an n-out-of-n sample could be more narrow, and thus under estimate the coverage. The solution proposed is to use an m-out-of-n bootstrap where $m < n$ [Hall and Miller, 2009]. In the study choosing an $m < n$ would be extremely difficult due to the unequal sample sizes, and with roughly 11% of the data containing less than 3 samples. Hall and Miller propose an empirical selection of m that is ideally not too small, as a small m can either not correct the problem and reduce your sample or make the problem worse. The problem arises where either the ideal m can be selected for the majority of the data, but drop at least 11% of the data, or conduct an n-out-of-n bootstrap and acknowledge the short coming.

6.4 Potential Future Research

The study hopefully lays the groundwork for better understanding graduate statistics departments. The most burdensome aspect of the study was collecting the individual faculty information. Now that this has been collected and built into a dataset, other research can either branch off by collecting additional information, or add these into existing ranking systems to improve existing rankings. There are multiple areas that could provide important information by conducting additional research.

First, further examine the impact that Ph.D. granting institutions have on rankings. For example, do some departments favor certain departments over another for hiring their tenure faculty? Does that have any correlation with research output? Second, expand the rankings to include more universities. Expanding the rankings would only allow for more coverage and allow people to further understand the structure of graduate statistics departments within the United States or perhaps globally. Lastly, find a way to generate a research statistic or metric that could be extrapolated so that there is less reliance on Google Scholar metrics. The main limitation of the study was reducing the sample size due to a lack of Google Scholar profile. Even just being able to find a way to incorporate ResearchGate and Google Scholar would give future research excellence coverage.

Appendix A U.S. News & World Report Rankings

Table 2: U.S. News & World Report statistics graduate program rankings

U.S. News and World Report Best Statistics Programs		
Rank	Name	Peer Assessment Score
1	Stanford University	4.9
2	University of California - Berkeley	4.8
3	Harvard University	4.6
3	University of Chicago	4.6
5	Carnegie Mellon University	4.4
5	Columbia University	4.4
7	Duke University	4.3
7	University of Michigan - Ann Arbor	4.3
7	University of Pennsylvania	4.3
7	University of Washington	4.3
11	North Carolina State University	4.1
11	University of North Carolina - Chapel Hill	4.1
13	Cornell University	4.0
13	Texas A&M University - College Station	4.0
13	University of California - Davis	4.0
13	University of Minnesota - Twin Cities	4.0
13	University of - Wisconsin Madison	4.0
13	Yale University	4.0
19	Iowa State University	3.9

19	Pennsylvania State University	3.9
19	University of California - Los Angeles	3.9
22	Purdue University - West Lafayette	3.8
22	University of Illinois - Urbana Champaign	3.8
24	Ohio State University	3.7
24	Rutgers University - New Brunswick	3.7
26	University of Florida	3.6
27	University of California Irvine	3.5
27	University of Texas Austin	3.5
29	Rice University	3.4
30	Colorado State University	3.3
30	Florida State University	3.3
30	Michigan State University	3.3
30	University of Connecticut	3.3
30	University of Iowa	3.3
35	University of Georgia	3.2
35	University of Pittsburgh	3.2
37	New York University	3.1
37	Northwestern University	3.1
37	University of Missouri - Columbia	3.1
37	Virginia Tech	3.1
41	Boston University	3.0
41	George Washington University	3.0
41	University of California - San Diego	3.0
44	Temple University	2.9
44	University of California - Riverside	2.9
44	University of California - Santa Barbara	2.9
44	University of California - Santa Cruz	2.9
44	University of Virginia	2.9
49	Arizona State University	2.8

49	University of Rochester	2.8
51	University of Illinois - Chicago	2.7
51	University of Massachusetts - Amherst	2.7
51	University of South Carolina	2.7
54	George Mason University	2.6
54	Indiana University - Bloomington	2.6
54	Oregon State University	2.6
54	SMU	2.6
54	University of Arizona	2.6
54	University of Maryland - Baltimore County	2.6
54	Washington University in St. Louis	2.6
61	Clemson University	2.5
61	University of Notre Dame	2.5
63	Case Western Reserve University	2.4
63	University of Cincinnati	2.4
63	University of Kentucky	2.4
66	Baylor University	2.3
66	Kansas State University	2.3
66	University of Nebraska	2.3
66	University of Texas - Dallas	2.3
70	University of Colorado - Denver	2.2
70	University of North Carolina - Charlotte	2.2
72	Binghamton University - SUNY	2.1
72	Bowling Green State University	2.1
72	Lehigh University	2.1
72	University of Cenral Florida	2.1
72	University of New Mexico	2.1
72	Washington State University	2.1
72	Worcester Polytechnic Institute	2.1
79	Auburn University	2.0

79	University of Texas - San Antonio	2.0
79	Utah State University	2.0
82	Montana State University	1.9
82	New Jersey Institute of Technology	1.9
82	Northern Illinois University	1.9
82	University of Alabama	1.9
86	Marquette University	1.8
86	Oklahoma State University	1.8
86	Portland State University	1.8
86	University of Arkansas	1.8
86	University of North Carolina - Greensboro	1.8
86	University of South Florida	1.8
92	Central Michigan University	1.7
92	Old Dominion University	1.7
92	University of Nevada - Reno	1.7
96	South Dakota State University	1.6
96	University of Toledo	1.6
96	Western Michigan University	1.6
99	North Dakota State University	1.5
99	Oakland University	1.5
99	University of Northern Colorado	1.5

Appendix B Research Output Raw Data Ranks

Table 3: Research output ranks by the three Google Scholar metrics and overall research rank.

Research Output Ranks				
Department Name	h-index rank	i10-index rank	cites rank	Research rank
Alabama	62.00	59.00	68.00	63.00
Auburn	79.00	79.50	59.00	72.50
AzSU	33.00	34.00	29.00	32.00
Baylor	77.00	76.00	30.00	61.00
BGSU	90.50	77.00	91.00	86.17
Binghmtn	60.00	58.00	67.00	61.67
BostU	53.00	53.00	47.00	51.00
CentMich	80.00	82.00	70.00	77.33
Chic	8.00	8.00	4.00	6.67
Clemson	75.00	73.00	74.00	74.00
CMU	17.00	30.00	14.00	20.33
Columbia	9.00	10.00	17.00	12.00
Cornell	15.00	15.00	22.00	17.33
CSU	30.00	36.00	26.00	30.67
Duke	4.00	7.00	8.00	6.33
FSU	22.00	20.00	20.00	20.67
GMason	73.00	60.00	77.00	70.00
GWU	50.00	47.00	52.00	49.67
Harvard	5.00	5.00	7.00	5.67

Iowa	71.00	65.00	81.00	72.33
ISU	40.00	42.00	51.00	44.33
IU	26.00	83.00	57.00	55.33
KSU	93.00	92.00	71.00	85.33
Marq	68.00	72.00	90.00	76.67
MichTech	64.00	81.00	45.00	63.33
Minn	34.00	38.00	42.00	38.00
MontSU	84.00	79.50	78.00	80.50
MSU	42.00	40.00	56.00	46.00
NCSU	28.00	22.00	41.00	30.33
NDSU	96.00	96.00	55.00	82.33
NJIT	78.00	95.00	94.00	89.00
NorthW	13.00	16.00	5.00	11.33
NotreDame	36.00	28.00	24.00	29.33
NYU	52.00	54.00	62.00	56.00
Oaklnd	88.00	84.00	87.00	86.33
OkSt	76.00	67.00	96.00	79.67
OldDom	95.00	93.00	89.00	92.33
OregSt	66.00	75.00	72.00	71.00
OSU	86.00	78.00	50.00	71.33
Penn	3.00	1.00	6.00	3.33
Pitt	24.00	29.00	39.00	30.67
PortSt	63.00	66.00	86.00	71.67
PSU	19.00	24.00	23.00	22.00
Purdue	58.00	61.00	58.00	59.00
Rice	31.00	31.00	27.00	29.67
Rutgers	44.00	50.00	44.00	46.00
SDSU	81.00	87.00	84.00	84.00
SMU	90.50	90.00	95.00	91.83
Stan	1.00	3.00	2.00	2.00

TAMU	21.00	21.00	36.00	26.00
Temp	72.00	74.00	43.00	63.00
Toldeo	20.00	14.00	9.00	14.33
UArk	82.00	86.00	25.00	64.33
UAz	27.00	27.00	28.00	27.33
UCBerk	2.00	2.00	1.00	1.67
UCDavis	32.00	33.00	21.00	28.67
UCDenver	25.00	26.00	18.00	23.00
UCF	74.00	71.00	92.00	79.00
UCin	56.00	57.00	54.00	55.67
UCLA	10.00	9.00	10.00	9.67
UConn	57.00	48.00	31.00	45.33
UCIrv	18.00	18.00	32.00	22.67
UCRiv	37.00	25.00	48.00	36.67
UCSB	51.00	45.00	75.00	57.00
UCSC	43.00	37.00	65.00	48.33
UCSD	14.00	11.00	19.00	14.67
UFla	59.00	56.00	63.00	59.33
UGA	69.00	64.00	66.00	66.33
UIC	67.00	55.00	80.00	67.33
UIll	23.00	23.00	34.00	26.67
UKent	45.00	51.00	38.00	44.67
UMass	61.00	68.00	16.00	48.33
UMBC	89.00	91.00	82.00	87.33
UMich	12.00	13.00	13.00	12.67
UMiss	55.00	62.00	53.00	56.67
UNC	46.00	43.00	76.00	55.00
UNCChar	92.00	89.00	88.00	89.67
UNCGreen	87.00	85.00	93.00	88.33
UNL	39.00	39.00	64.00	47.33

UNM	94.00	94.00	79.00	89.00
UNReno	47.00	52.00	49.00	49.33
URoch	35.00	32.00	33.00	33.33
USCar	70.00	70.00	73.00	71.00
USFl	85.00	69.00	85.00	79.67
UtahSt	29.00	17.00	40.00	28.67
UTAust	7.00	6.00	15.00	9.33
UTDall	65.00	63.00	60.00	62.67
UTSA	54.00	46.00	69.00	56.33
UVA	38.00	35.00	35.00	36.00
UWash	6.00	4.00	3.00	4.33
Uwisc	16.00	19.00	12.00	15.67
VaTech	41.00	41.00	46.00	42.67
WashSt	83.00	88.00	83.00	84.67
WorchPI	49.00	44.00	61.00	51.33
WuSL	48.00	49.00	37.00	44.67
Yale	11.00	12.00	11.00	11.33

Appendix C Research Rank Outputs Raw Data Continued

Table 4: Research Rank and U.S. News & World Report Rank and RMSE

Research Output Ranks			
Department Name	Research rank	USNWR rank	RMSE
Alabama	63.00	82	19.36
Auburn	72.50	79	11.55
AzSU	32.00	49	17.13
Baylor	61.00	66	22.48
BGSU	86.17	72	15.58
Binghmtn	61.67	72	11.03
BostU	51.00	41	10.39
CentMich	77.33	92	15.57
Chic	6.67	3	4.12
Clemson	74.00	61	13.03
CMU	20.33	5	16.83
Columbia	12.00	5	7.85
Cornell	17.33	13	5.44
CSU	30.67	30	4.16
Duke	6.33	7	1.83
FSU	20.67	30	9.38
GMason	70.00	54	17.56
GWU	49.67	41	8.90
Harvard	5.67	3	2.82

Iowa	72.33	30	42.84
ISU	44.33	19	25.78
IU	55.33	54	23.33
KSU	85.33	66	21.83
Marq	76.67	86	13.36
MichTech	63.33	92	32.21
Minn	38.00	13	25.21
MontSU	80.50	82	2.95
MSU	46.00	30	17.51
NCSU	30.33	11	20.89
NDSU	82.33	99	25.52
NJIT	89.00	82	10.47
NorthW	11.33	37	26.08
NotreDame	29.33	61	32.05
NYU	56.00	37	19.48
Oaklnd	86.33	99	12.78
OCIrv	22.67	27	7.89
OkSt	79.67	86	13.67
OldDom	92.33	92	2.51
OregSt	71.00	54	17.40
OSU	71.33	24	49.78
Penn	3.33	7	4.20
Pitt	30.67	35	7.59
PortSt	71.67	86	17.59
PSU	22.00	19	3.69
Purdue	59.00	22	37.02
Rice	29.67	29	2.00
Rutgers	46.00	24	22.18
SDSU	84.00	96	12.24
SMU	91.83	54	37.90

Stan	2.00	1	1.29
TAMU	26.00	13	14.79
Temp	63.00	44	23.69
Toldeo	14.33	96	81.79
UArk	64.33	86	35.29
UAz	27.33	54	26.67
UCBerk	1.67	2	0.57
UCDavis	28.67	13	16.58
UCDenver	23.00	70	47.13
UCF	79.00	72	11.61
UCin	55.67	63	7.43
UCLA	9.67	19	9.34
UConn	45.33	30	18.74
UCRiv	36.67	44	11.91
UCSB	57.00	44	18.35
UCSC	48.33	44	12.79
UCSD	14.67	41	26.53
UFla	59.33	26	33.45
UGA	66.33	35	31.40
UIC	67.33	51	19.26
UIll	26.67	22	6.97
UKent	44.67	63	19.08
UMass	48.33	51	23.19
UMBC	87.33	54	33.55
UMich	12.67	7	5.68
UMiss	56.67	37	20.04
UNC	55.00	11	46.45
UNCChar	89.67	70	19.73
UNCGreen	88.33	86	4.12
UNL	47.33	82	36.61

UNM	89.00	72	18.41
UNReno	49.33	92	42.71
URoch	33.33	49	15.71
USCar	71.00	51	20.04
USFl	79.67	86	9.94
UtahSt	28.67	79	51.20
UTAust	9.33	27	18.11
UTDall	62.67	66	3.91
UTSA	56.33	79	24.58
UVA	36.00	44	8.12
UWash	4.33	7	2.94
Uwisc	15.67	13	3.91
VaTech	42.67	37	6.13
WashSt	84.67	72	12.88
WorchPI	51.33	72	21.86
WuSL	44.67	54	10.80
Yale	11.33	13	1.73

Appendix D Google Scholar Metrics

Table 5: Google Scholar z-score median metrics by department

Google Scholar Metrics by Department			
Department Name	Median h-index z-score	Median i10-index z-score	Median total cites z-score
Alabama	-0.50	-0.52	-0.44
Auburn	-0.76	-0.72	-0.41
AzSU	-0.16	-0.24	-0.26
Baylor	-0.74	-0.68	-0.26
BGSU	-0.90	-0.71	-0.63
Binghmtn	-0.50	-0.50	-0.44
BostU	-0.42	-0.45	-0.38
CentMich	-0.76	-0.73	-0.45
Chic	0.50	0.51	0.34
Clemson	-0.73	-0.63	-0.47
CMU	0.06	-0.15	-0.05
Columbia	0.47	0.34	-0.17
Cornell	0.15	0.07	-0.20
CSU	-0.11	-0.27	-0.23
Duke	0.91	0.55	0.27
FSU	-0.06	-0.05	-0.20
GMason	-0.70	-0.52	-0.49
GWU	-0.41	-0.41	-0.39
Harvard	0.74	0.69	0.31

Iowa	-0.64	-0.56	-0.52
ISU	-0.33	-0.33	-0.39
IU	-0.09	-0.74	-0.40
KSU	-0.98	-0.81	-0.46
Marq	-0.60	-0.63	-0.62
MichTech	-0.56	-0.73	-0.37
Minn	-0.16	-0.29	-0.34
MontSU	-0.78	-0.72	-0.50
MSU	-0.34	-0.30	-0.40
NCSU	-0.09	-0.08	-0.31
NDSU	-1.06	-0.99	-0.40
NJIT	-0.75	-0.95	-0.65
NorthW	0.25	0.04	0.34
NotreDame	-0.17	-0.14	-0.21
NYU	-0.42	-0.45	-0.42
Oaklnd	-0.83	-0.75	-0.57
OCIrv	0.04	0.02	-0.28
OkSt	-0.73	-0.56	-0.68
OldDom	-1.03	-0.82	-0.59
OregSt	-0.59	-0.66	-0.46
OSU	-0.83	-0.72	-0.39
Penn	1.34	1.12	0.34
Pitt	-0.07	-0.15	-0.31
PortSt	-0.53	-0.56	-0.57
PSU	-0.02	-0.11	-0.21
Purdue	-0.48	-0.53	-0.41
Rice	-0.13	-0.16	-0.23
Rutgers	-0.36	-0.44	-0.36
SDSU	-0.76	-0.78	-0.55
SMU	-0.90	-0.80	-0.65

Stan	1.40	0.98	0.72
TAMU	-0.03	-0.07	-0.29
Temp	-0.66	-0.64	-0.35
Toldeo	-0.03	0.08	0.19
UArk	-0.76	-0.77	-0.22
UAz	-0.09	-0.12	-0.24
UCBerk	1.36	1.04	1.13
UCDavis	-0.15	-0.23	-0.20
UCDenver	-0.07	-0.12	-0.19
UCF	-0.72	-0.59	-0.63
UCin	-0.46	-0.50	-0.40
UCLA	0.44	0.39	0.15
UConn	-0.47	-0.42	-0.27
UCRiv	-0.17	-0.12	-0.38
UCSB	-0.41	-0.39	-0.48
UCSC	-0.35	-0.29	-0.43
UCSD	0.24	0.22	-0.20
UFla	-0.48	-0.48	-0.43
UGA	-0.62	-0.56	-0.44
UIC	-0.60	-0.47	-0.50
UIll	-0.06	-0.11	-0.28
UKent	-0.37	-0.44	-0.31
UMass	-0.50	-0.57	-0.16
UMBC	-0.87	-0.81	-0.54
UMich	0.34	0.12	-0.02
UMiss	-0.45	-0.54	-0.40
UNC	-0.37	-0.34	-0.48
UNCChar	-0.93	-0.80	-0.59
UNCGreen	-0.83	-0.75	-0.65
UNL	-0.31	-0.29	-0.43

UNM	-1.00	-0.90	-0.50
UNReno	-0.38	-0.44	-0.39
URoch	-0.17	-0.18	-0.28
USCar	-0.64	-0.59	-0.46
USFl	-0.80	-0.58	-0.55
UtahSt	-0.10	0.03	-0.31
UTAust	0.51	0.66	-0.09
UTDall	-0.58	-0.55	-0.42
UTSA	-0.44	-0.39	-0.45
UVA	-0.21	-0.24	-0.29
UWash	0.72	0.89	0.63
Uwisc	0.14	0.01	0.03
VaTech	-0.33	-0.32	-0.37
WashSt	-0.77	-0.79	-0.54
WorchPI	-0.40	-0.38	-0.42
WuSL	-0.40	-0.43	-0.30
Yale	0.42	0.18	0.08

Appendix E Bootstrap Confidence Intervals Data

Table 6: Google Scholar bootstrapped quantile confidence intervals.

95% Bootstrap Confidence Interval Ranks				
Department Name	Total Cites	h-index	i10-index	Research Rank
Alabama	(37.00, 86.00)	(25.00, 88.00)	(22.00, 87.00)	(33.66, 85.67)
Auburn	(48.00, 85.00)	(57.00, 92.025)	(43.00, 95.00)	(51.67, 88.67)
AzSU	(8.00, 69.00)	(6.00, 78.00)	(11.98, 77.00)	(7.99, 68.67)
Baylor	(1.00, 92.00)	(32.00, 95.00)	(25.00, 93.00)	(32.00, 87.67)
BGSU	(69.00, 95.00)	(60.50, 96.00)	(58.50, 89.00)	(63.67, 91.67)
Binghmtn	(26.00, 78.00)	(50.00, 81.00)	(45.00, 69.00)	(54.33, 63.67)
BostU	(10.00, 87.00)	(4.00, 73.50)	(6.00, 73.00)	(7.33, 76.00)
CentMich	(46.98, 83.00)	(56.00, 95.00)	(60.00, 96.00)	(62.00, 84.67)
Chic	(2.00, 15.00)	(2.99, 21.00)	(4.00, 16.02)	(3.67, 16.67)
Clemson	(38.00, 87.00)	(55.00, 87.00)	(50.98, 78.00)	(50.33, 81.67)
CMU	(7.00, 33.00)	(10.00, 52.00)	(12.00, 57.01)	(11.33, 45.01)
Columbia	(6.00, 35.00)	(5.00, 39.00)	(4.00, 25.00)	(6.33, 32.00)
Cornell	(12.00, 72.00)	(8.50, 54.00)	(8.00, 66.00)	(10.50, 62.00)
CSU	(12.00, 66.00)	(11.00, 71.00)	(13.00, 79.00)	(14.00, 72.00)
Duke	(3.00, 17.00)	(1.00, 20.00)	(1.00, 19.00)	(2.33, 18.33)
FSU	(17.00, 37.00)	(7.00, 46.02)	(8.00, 47.00)	(11.99, 42.34)
GMason	(55.00, 94.00)	(46.98, 88.00)	(38.00, 90.00)	(47.00, 90.01)
GWU	(29.00, 84.00)	(37.00, 65.00)	(19.00, 66.00)	(32.33, 67.35)

Harvard	(1.00, 19.00)	(1.00, 35.00)	(1.00, 32.00)	(1.33, 26.33)
Iowa	(36.00, 85.00)	(38.00, 79.17)	(39.49, 87.00)	(38.50, 83.39)
ISU	(28.98, 70.00)	(29.00, 69.01)	(23.98, 70.02)	(28.00, 66.67)
IU	(3.00, 82.02)	(7.00, 93.00)	(8.00, 94.00)	(7.67, 90.00)
KSU	(18.00, 94.00)	(11.98, 96.00)	(12.00, 95.00)	(14.00, 93.01)
Marq	(64.00, 96.00)	(52.99, 77.01)	(52.00, 86.00)	(56.99, 86.00)
MichTech	(41.00, 89.00)	(24.49, 73.00)	(14.00, 95.00)	(28.33, 78.00)
Minn	(25.98, 58.00)	(18.98, 47.02)	(22.00, 52.00)	(26.99, 50.00)
MontSU	(40.00, 88.00)	(35.98, 96.00)	(41.98, 94.00)	(39.66, 92.33)
MSU	(32.00, 74.00)	(21.00, 60.00)	(15.00, 56.00)	(25.32, 60.17)
NCSU	(20.00, 64.00)	(15.99, 43.00)	(12.00, 57.00)	(17.16, 51.00)
NDSU	(38.98, 73.00)	(67.00, 96.00)	(70.00, 96.00)	(69.00, 83.33)
NJIT	(66.00, 96.00)	(55.00, 95.00)	(37.00, 96.00)	(61.33, 95.33)
NorthW	(1.00, 26.02)	(3.00, 53.02)	(4.00, 61.02)	(2.67, 46.67)
NotreDame	(17.98, 82.00)	(14.00, 71.00)	(12.00, 83.00)	(16.99, 79.01)
NYU	(35.00, 95.00)	(27.98, 94.00)	(27.00, 96.00)	(30.99, 94.34)
Oaklnd	(40.00, 96.00)	(79.00, 89.02)	(75.50, 88.00)	(69.00, 88.00)
OCIrv	(14.00, 41.00)	(8.00, 42.00)	(10.00, 43.00)	(12.00, 40.33)
OkSt	(88.00, 96.00)	(62.50, 88.00)	(58.00, 71.00)	(71.67, 82.67)
OldDom	(74.00, 96.00)	(64.00, 96.00)	(75.97, 95.00)	(72.33, 94.67)
OregSt	(21.00, 75.00)	(35.00, 86.00)	(37.00, 90.00)	(33.33, 81.51)
OSU	(28.00, 68.00)	(47.78, 95.00)	(33.00, 94.00)	(35.66, 83.00)
Penn	(2.00, 16.00)	(1.00, 8.0)0	(1.00, 8.00)	(1.67, 9.33)
Pitt	(14.00, 93.00)	(16.98, 92.00)	(20.00, 96.00)	(20.00, 92.67)
PortSt	(30.00, 89.00)	(37.00, 79.00)	(30.00, 92.00)	(51.33, 84.00)
PSU	(8.00, 58.00)	(5.98, 52.02)	(8.00, 58.00)	(9.33, 51.34)
Purdue	(33.00, 74.00)	(40.00, 76.00)	(43.00, 73.00)	(40.33, 72.67)
Rice	(18.00, 96.00)	(8.00, 95.00)	(2.00, 95.00)	(12.83, 95.33)
Rutgers	(31.98, 84.00)	(24.00, 63.00)	(23.00, 66.01)	(28.99, 64.17)
SDSU	(75.00, 92.00)	(74.00, 93.51)	(69.00, 91.00)	(75.33, 88.33)

SMU	(34.00, 95.00)	(13.00, 96.00)	(13.00, 96.00)	(20.00, 95.00)
Stan	(1.00, 14.00)	(1.00, 13.00)	(1.00, 13.00)	(1.00, 12.01)
TAMU	(18.00, 66.00)	(17.49, 45.51)	(18.00, 56.00)	(19.33, 51.85)
Temp	(26.00, 77.00)	(27.98, 87.00)	(38.98, 86.00)	(34.66, 78.67)
Toldeo	(2.00, 90.00)	(4.00, 84.00)	(3.00, 87.00)	(3.33, 86.33)
UArk	(17.00, 42.00)	(58.49, 92.00)	(66.00, 93.00)	(47.50, 74.67)
UAz	(5.00, 91.00)	(10.00, 85.00)	(7.00, 89.00)	(11.00, 87.00)
UCBerk	(1.00, 8.00)	(1.00, 8.00)	(1.00, 13.02)	(1.00, 9.00)
UCDavis	(12.00, 49.00)	(8.00, 58.00)	(9.00, 63.00)	(10.67, 54.67)
UCDenver	(12.00, 84.02)	(7.00, 85.50)	(6.00, 87.00)	(9.83, 85.67)
UCF	(28.00, 95.00)	(9.00, 86.51)	(7.00, 88.00)	(15.00, 89.34)
UCin	(42.00, 92.00)	(38.00, 93.00)	(40.00, 94.00)	(42.83, 91.00)
UCLA	(2.00, 28.00)	(4.00, 22.00)	(5.00, 34.00)	(4.00, 26.33)
UConn	(22.00, 56.00)	(17.98, 75.00)	(9.00, 81.00)	(17.00, 65.33)
UCRiv	(17.00, 69.00)	(17.00, 57.00)	(15.98, 54.00)	(21.67, 59.33)
UCSB	(24.98, 87.00)	(32.00, 82.00)	(32.00, 84.00)	(32.65, 83.67)
UCSC	(4.00, 81.00)	(12.00, 90.00)	(13.00, 93.00)	(15.33, 82.33)
UCSD	(17.00, 75.00)	(11.00, 47.00)	(9.00, 41.00)	(13.33, 54.01)
UFla	(37.00, 87.00)	(34.98, 67.00)	(32.98, 65.51)	(40.31, 70.67)
UGA	(33.98, 94.00)	(43.00, 84.00)	(36.00, 87.02)	(43.33, 82.01)
UIC	(45.00, 81.00)	(15.00, 77.00)	(9.00, 81.00)	(25.00, 76.84)
UIll	(19.00, 54.00)	(19.00, 40.00)	(19.00, 39.00)	(20.32, 41.33)
UKent	(21.00, 52.00)	(14.00, 91.50)	(12.00, 92.01)	(17.99, 77.00)
UMass	(7.00, 64.02)	(12.50, 89.00)	(19.00, 91.00)	(17.33, 80.33)
UMBC	(70.00, 95.00)	(48.50, 95.00)	(35.49, 96.00)	(55.00, 94.67)
UMich	(6.00, 22.00)	(4.00, 24.00)	(5.00, 28.00)	(6.33, 23.33)
UMiss	(42.98, 84.00)	(29.00, 80.00)	(32.00, 78.02)	(36.50, 78.67)
UNC	(25.98, 85.00)	(27.00, 77.02)	(26.00, 78.00)	(28.16, 78.00)
UNCChar	(35.00, 95.00)	(61.00, 95.00)	(69.00, 94.00)	(55.66, 92.33)
UNCGreen	(74.00, 96.00)	(68.00, 92.00)	(60.50, 94.00)	(74.33, 89.00)

UNL	(13.98, 95.00)	(15.00, 87.00)	(16.00, 85.50)	(15.33, 89.00)
UNM	(43.98, 89.00)	(59.00, 96.00)	(68.00, 96.00)	(63.00, 93.33)
UNReno	(38.00, 88.00)	(37.00, 56.50)	(45.00, 61.51)	(42.33, 61.67)
URoch	(12.98, 70.00)	(14.00, 50.00)	(13.00, 49.50)	(14.67, 55.01)
USCar	(37.00, 91.00)	(58.00, 89.00)	(55.00, 88.02)	(54.67, 86.67)
USFl	(67.00, 96.00)	(38.00, 96.00)	(34.00, 96.00)	(47.67, 96.00)
UtahSt	(17.00, 95.00)	(6.00, 87.02)	(5.00, 91.00)	(9.99, 76.67)
UTAust	(8.00, 33.00)	(3.00, 42.00)	(1.00, 40.00)	(4.67, 37.34)
UTDall	(42.98, 72.00)	(46.00, 92.00)	(30.00, 90.00)	(44.33, 82.00)
UTSA	(15.00, 75.02)	(24.00, 88.00)	(13.00, 88.00)	(24.33, 81.00)
UVA	(23.00, 83.00)	(17.00, 95.00)	(23.00, 92.00)	(24.00, 89.00)
UWash	(1.00, 20.00)	(1.00, 20.00)	(1.00, 25.02)	(1.00, 20.33)
Uwisc	(7.00, 25.00)	(5.00, 33.00)	(7.00, 44.00)	(7.33, 31.00)
VaTech	(20.00, 64.02)	(19.00, 57.02)	(26.00, 57.00)	(22.99, 57.50)
WashSt	(40.00, 94.02)	(41.00, 95.01)	(45.98, 94.00)	(52.83, 90.00)
WorchPI	(29.00, 90.00)	(29.00, 88.00)	(30.00, 81.00)	(31.33, 84.67)
WuSL	(32.00, 96.00)	(23.00, 95.00)	(17.00, 95.00)	(26.00, 95.33)
Yale	(1.00, 18.00)	(4.00, 19.00)	(4.00, 26.00)	(3.67, 18.33)

Bibliography

[Barro and Lee, 2013] Barro, R. J. and Lee, J.-W. (2013). A new data set of education attainment in the world 1950-2010. *Journal of Development Economics*, 41:184–198.

[Bowman and Bastedo, 2010] Bowman, N. and Bastedo, M. N. (2010). U.s. news world report college rankings: Modeling institutional effects on organizational reputation. *American Journal of Education*, 116:163–183.

[Clarke, 2002] Clarke, M. (2002). Quantifying quality: What can the u.s. news and world report rankings tell us about the quality of higher education? *Education Policy Analysis Archives*, 10.

[Dichev, 2001] Dichev, I. (2001). News or noise? estimating the noise in the u.s. news university rankings. 42:237–266.

[Hall and Miller, 2009] Hall, P. and Miller, H. (2009). Using the bootstrap to quantify the authority of an empirical ranking. *The Annals of Statistics*, 37:3929–3959.

[Kreuter et al., 2008] Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social desirability bias in cati, ivr, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72:847–865.

[Lee and Lee, 2016] Lee, J.-W. and Lee, H. (2016). Human capital in the long run. *Journal of Development Economics*, 122:147–169.

[Machung, 1998] Machung, A. (1998). Playing the ranking game. *Change: The Magazine of Higher Learning*, 30:12–16.

[Meredith, 2004] Meredith, M. (2004). Why do universities compete in the ratings game? an empirical analysis of the effects of the u.s. news and world report college rankings. *Research in Higher Education*, 45:443–461.

[Merton, 1968] Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159:56–63.

[Monks and Ehrenberg, 1999] Monks, J. and Ehrenberg, R. G. (1999). The impact of u.s. news world report college rankings on admissions outcomes and pricing policies at selective private institutions. *National Bureau of Economic Research*.

[Myers and Robe, 2009] Myers, L. and Robe, J. (2009). College rankings: History, criticism and reform. *Center for College Affordability and Productivity*.

[Safón and Docampo, 2020] Safón, V. and Docampo, D. (2020). Analyzing the impact of reputational bias on global university rankings based on objective research performance data: the case of the shanghai ranking (awru). *Scientometrics*, 125:2199–2227.

[Webster, 2001] Webster, T. J. (2001). A principal component analysis of the u.s. news world report tier rankings of colleges and universities. *Economics of Education Review*, 20:235–244.