Scholarly Influence and the Shaping of International Relations Debates

An International Studies Quarterly Online symposium

Daniel Nexon
Katie Paulson-Smith
Michael Tierney
Cullen Hendrix
Jelena Subotic
Jeff Colgan

DeRaismes Combes, Managing Editor

Published Online, 1 August 2017 v1.0

INTRODUCTION

Daniel Nexon
Georgetown University

Back in 2011, I characterized the field as having moved from the "paradigm wars" to the "war on paradigms." (Nexon 2011). That same year, David Lake proclaimed that "isms" are "evil" in the pages of International Studies Quarterly. (Lake 2011). Indeed, by 2015 the European Journal of International Relations was debating the "End of IR Theory." (Nexon, 2015). But while international-relations scholars variously celebrated, rent their garments, or yawned, the fact is that we don't know all that much about the actual state of theory and theorizing in the field. In his recent ISQ piece, "Where is International Relations Going? Evidence from Graduate Training," (2016) Jeff Colgan seeks to shed more light on the present and future of the field. In doing so, he takes on many trends, including the increasing reliance on journal-oriented metrics to allocate status and prestige. (Hendrix 2015, Mugge 2015, Nexon and Jackson 2015).

And in this symposium, four scholars respond to his claims in three pieces. Katie Paulson-Smith and Michael J. Tierney focus on Colgan's "Teaching Influence Score (TIS)" and how it compares to other metrics of influence. Cullen Hendrix offers a critical assessment of TIS. Jelena Subotic finds TIS problematic as well, but for reasons that implicate all metric-based assessments of scholarly importance. Finally, Colgan responds to his interlocutors.

How to Count What Counts: TIS the Season for Syllabi Metrics?

Katie Paulson-Smith and Michael J. Tierney
University of Madison-Wisconsin

Jeff Colgan's recent paper advances the study of the international relations (IR) discipline in three ways. First, he empirically explores a series of prominent and untested claims about the direction of the field (Mearsheimer and Walt 2013; Aggarwal 2010). Second, he provides a new method for measuring the impact of published scholarship. Finally, he generates a series of plausible and interesting claims about the field -- some of which he tests and some that remain to be explored. This last feature is one of the most valuable parts of Colgan's contribution. Despite the fact that this is an empirical paper, the idea-to-word ratio is very high, and the paper produces new ideas that serve as an inviting springboard for wild speculation (and future research) for the rest of us.

The editors of *ISQ* have invited several response essays and we expect these will be full of empirical, conceptual, and normative critiques. We will join the fray briefly near the end of this essay, mostly with quibbles rather than foundational critiques, but will use the bulk of our essay to present data from the Teaching, Research, and International Policy (TRIP) faculty surveys that speak to a number of Colgan's timely questions. We show that there is disagreement among IR scholars in the United States about whether various citation metrics are good measures of scholarly impact, or whether and how they should be used in the tenure and promotion process. We provide evidence that suggests IR scholars assess impact differently depending upon their gender, academic rank, methodology, analytic approach, epistemology, and type of institution. We also find that the way faculty organize the IR field seminar is broadly consistent with Colgan's findings about the type of readings assigned in field seminar syllabi. Like Colgan, we hope these results generate more introspection, conversation, and research on disciplinary practices within IR.

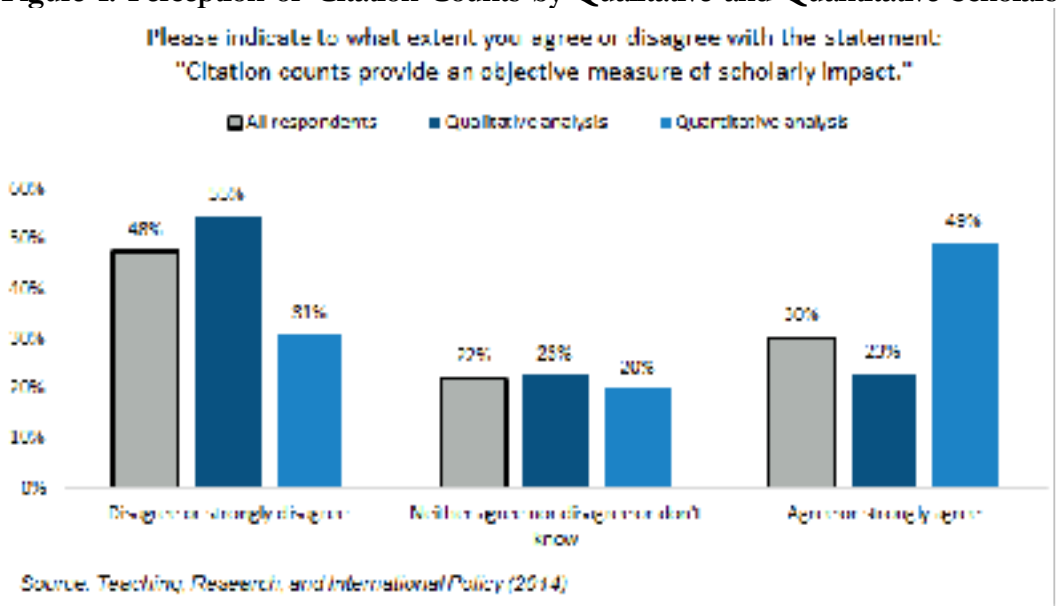
Citation Counts Don't Count Everything

Colgan joins a growing chorus of scholars who observe that the academy is increasingly obsessed with measuring and demonstrating the impact of scholarship. Individual scholars seek to demonstrate the impact of their work in order to achieve tenure and promotion, departments to impress administrators or prospective students, and universities to maintain rankings and thus resources. While various citation metrics (Web of Science and Google Scholar) and surveys (TRIP and Garand and Giles 2003) have been used to assess the impact of scholarship and ideas, there have been very few efforts to study impact by measuring which books, journals, or specific articles are included on course syllabi (Hagmann and Biersteker 2012). Presumably professors select research that they believe will be most useful in teaching the discipline to the next generation of IR scholars... or practitioners.

Colgan is skeptical of "various metrics based on citation counts," and uses this as one justification for creating the "Teaching Influence Score" (TIS). But is this skepticism (discussed here, here, and here) (Samuels 2013; Hendrix 2015; Nexon and Jackson 2015)

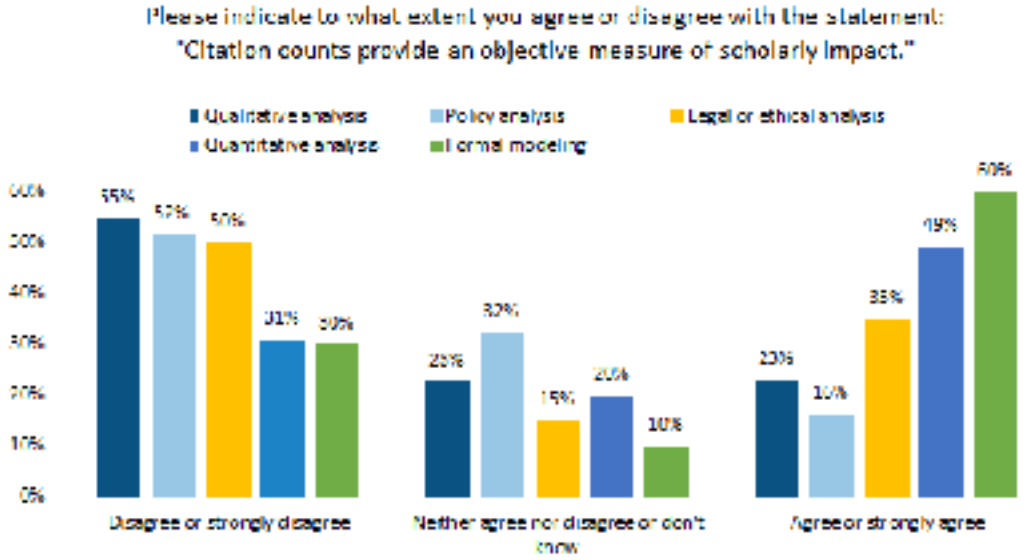
reflected in the views of all IR scholars? We use data from the 2014 TRIP Faculty Survey of IR scholars based at U.S. institutions to explore what types of scholars are likely to see various citation metrics as objective and/or useful in assessing impact. When asked whether "citation counts provide an objective measure of scholarly impact," about half disagreed while less than a third agreed. In a robust pattern (see Figure 1) that recurs in other questions below, we see greater enthusiasm for citation counts among quantitative scholars than those who employ qualitative methods.

Figure 1. Perception of Citation Counts by Qualitative and Quantitative Scholars



Source: *Teaching, Research, and International Policy (2014)*

Figure 2. Perception of Citation Counts by All Methods

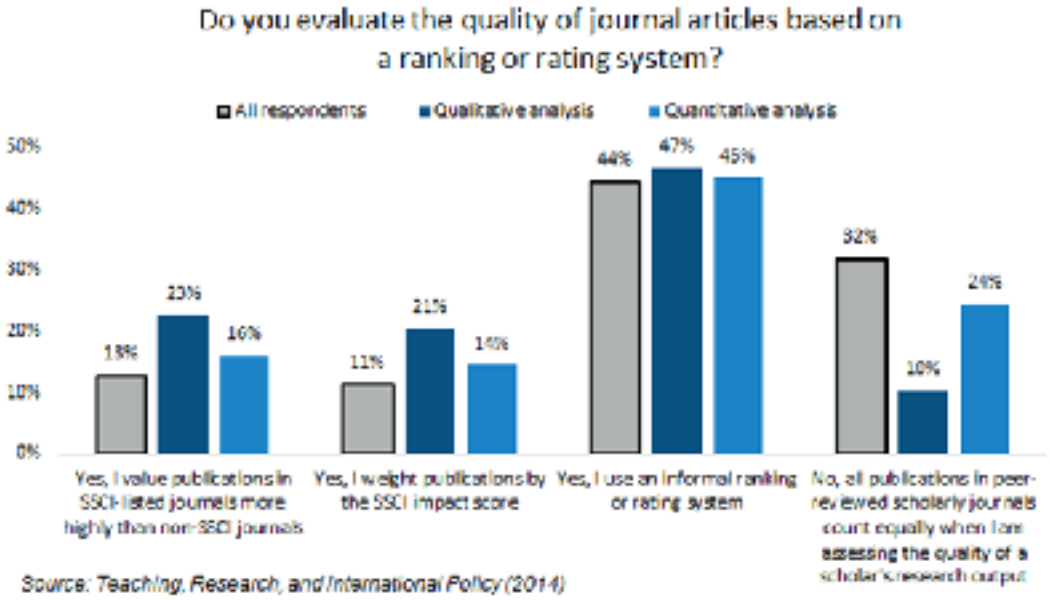


Source: *Teaching, Research, and International Policy (2014)*

While we do not have as many observations for scholars who use other types of methods, and thus cannot be as confident in the representativeness of their responses, a similar pattern emerges when we include policy analysis, legal/ethical analysis, and formal modeling as methods. Here we observe that quantitative scholars and formal modelers are more likely to agree with the premise that citation counts represent an objective measure of influence (see Figure 2). This finding persists despite the fact that articles using quantitative methods are systematically less likely to be cited, and citation rates decay more quickly than articles that Saideman (2015) codes as "grand theory" articles or non-formal IR theory articles.

Colgan cites Curry's (2012) discussion of Thomson Reuters SSCI impact factor when he claims: "if you use impact factors you are statistically illiterate." For those who accept Curry's premise, it may be distressing, or mildly amusing to learn that when you slice the data by methodology, scholars who employ quantitative methods were about twice as likely to report that they "weight publications by their SSCI impact score" as displayed in the second column of Figure 3 below.

Figure 3. Evaluating Journal Articles with Ranking or Rating System



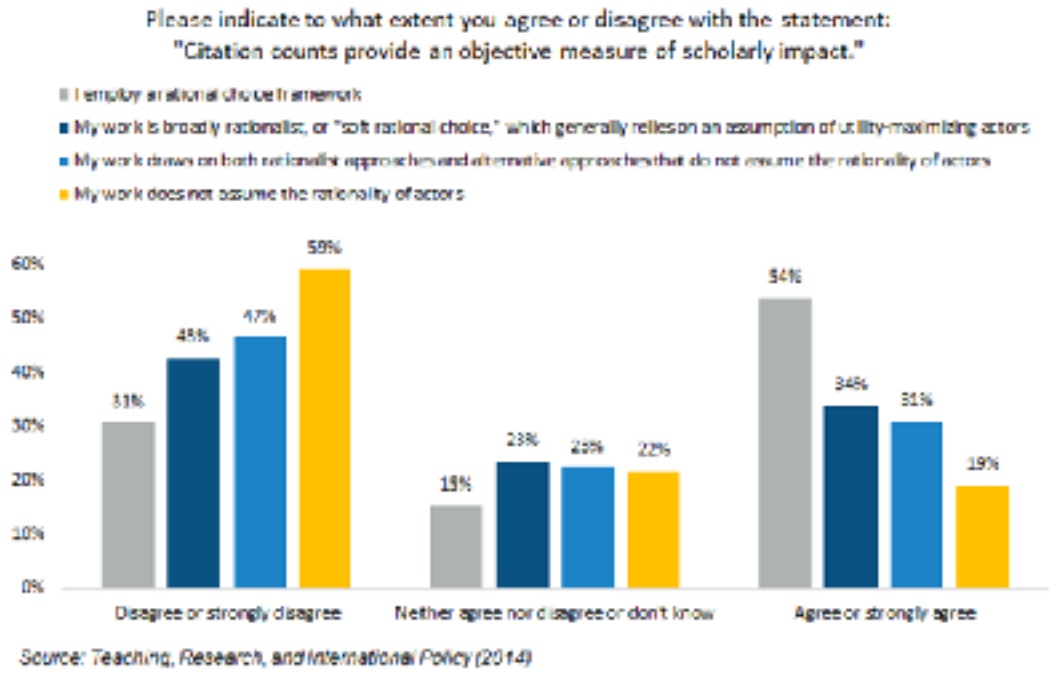
Source: *Teaching, Research, and International Policy (2014)*

In addition to methodology, we found that the analytic perspective one brings to the study of politics also influences one's faith in citation counts. Respondents who claim to employ a rational choice framework tend to believe that citation counts provide an objective measure of scholarly impact (See Figure 4), while those who "do not assume the rationality of actors" are half as likely to agree. Similarly, rationalists report that they are twice as likely as non-rationalists to think Google Scholar citation counts or the h-index is important (46 percent compared to 19 percent, respectively). While not shown in the figure below, we see similar results for epistemology, where self-described "positivists" are more likely to see citation counts as an objective measure of impact and to use them in their assessments for tenure and promotion.

We guessed that anecdotes about the use of various citation metrics in the tenure and promotion process likely reflect emerging practices among scholars at R-1 institutions,

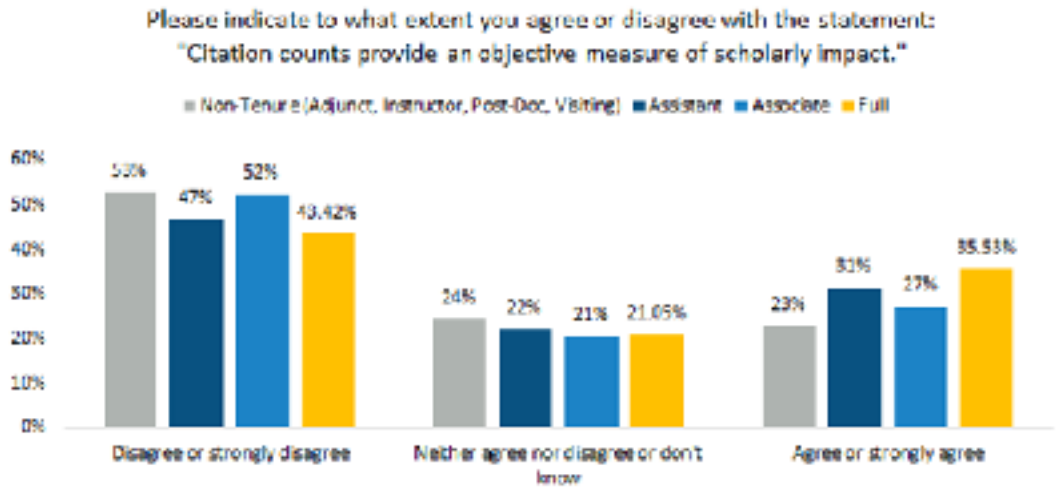
where the pressure to publish is most intense. However, scholars' views on this issue at research universities are almost the same as their colleagues' views at liberal arts colleges or comprehensive four-year institutions. Similarly, while one might expect tenure status to have a large effect on the perception of various citation metrics, the eyeball test (as illustrated in Figure 5) suggests only minor differences between scholars at different academic ranks with full professors being slightly more positive than their more junior colleagues.

Figure 4. Perception of Citation Counts by Analytic Framework



Source: Teaching, Research, and International Policy (2014)

Figure 5. Perception of Citation Counts by Academic Rank



Source: Teaching, Research, and International Policy (2014)

Finally, there has been much recent analysis and discussion about gender citation bias and even work in progress by Colgan about parallel underrepresentation of female scholars on syllabi (Maliniak, Powers, and Walter 2013; Colgan 2015; Østby et al. 2013; Wemheuer-Vogelaar 2013). So, readers may not be terribly surprised to learn that female IR scholars use citation metrics less frequently than their male counterparts and are more skeptical about their objectivity as measures of scholarly impact (Campbell and Desch 2013).

How to Improve TIS as a Measure of Impact

Most scholars tend to agree that citation measures do not and cannot capture everything we might want to measure; however, aside from individual judgments and external letters in the tenure and promotion process, it is one of the few metrics we have to measure scholarly impact. Colgan's new "Training Influence Score" (TIS) specifies another systematic and transparent measure of scholarly impact that, like citation metrics, reputational surveys, and other new efforts to measure scholarly impact offer partial and imperfect measures of the underlying concept. TIS can help to triangulate, to use a variety of measures meant to capture different types of scholarly impact that can be used in conjunction with more traditional assessments of research quality and/or impact. We provide a few "friendly amendments" that will make Colgan's measure even more useful if implemented, and then raise one fundamental limitation of TIS.

Scholars have argued here and here that one problem with citation metrics is that they capture both positive and negative citations, and perhaps we ought not be crediting scholars with negative citations (Rathbun 2011; Sabaratnam 2015). Colgan suggests that a similar effect could be at play in a work's inclusion on a syllabus. One of us (Tierney) assigns two of the discipline's most highly cited works (Huntington's *Clash of Civilizations* and Mearsheimer's "Back to the Future") of the past thirty years because they are so clearly written and so clearly wrong. These are wonderful pedagogical foils for the classroom. In the limited empirical research (Catalini et al. 2015) done on this topic, scholars have found that negative citations are actually more rare than most imagine them to be. We currently have a project underway with Lindsay Hundley that aims to measure this and other qualities of a citation to a given publication. If you think of a "negative citation" as one that expresses a negative sentiment about the quality of the cited work, preliminary work suggests that only about 2 percent of citations to the most cited IR books and articles are "negative" in this sense. If you broaden the definition to include any citation that disagrees with either the theory, methods, or conclusions of the cited source, we find "negative" citations to be around 20 percent of all citations.

Examining the Direction of IR

In terms of specific inferences that Colgan draws from his analysis, we focus on just one issue that likely has a direct impact on the findings. Colgan does a good job discussing the fact that IR is typically organized and taught as a subfield of political science at most U.S. universities, but that the publication outlets most valued by IR scholars are different from those valued by specialists of American and (to a lesser degree) Comparative Politics. However, some of his conclusions are more or less truer today than his results suggest. Recall, TIS draws upon a convenience sample of syllabi used between the years 2008 and 2013 with the majority of syllabi drawn from the most recent part of the time series in 2012 or 2013. Colgan compares various features of IR articles and journals that appear on syllabi

in this later period to a cross section of all articles/journals in the TRIP Journal Article Dataset from 1980 to 2006 (Maliniak and Powers 2015), which was the only data to which Colgan had access at the time he published this paper.

TRIP's most recent journal article data runs through 2012 and actually reaffirms/ strengthens some of Colgan's key findings. Specifically, the data between 2007 and 2012, that Colgan could not include in his analysis, contain a higher proportion of articles that employ quantitative methods and a lower proportion of "formal theory" and "analytic non-formal" articles (see Figure 6). These newer data actually strengthen Colgan's claim about the "gap" between what is published in top IR journals and what is taught in IR field seminars. However, newer data mildly weaken a second claim in the paper where Colgan demonstrates that the types of articles published in the top ranked IR journals (*IO, ISQ, IS*, and *WP*) are systematically different than IR articles published in the four "general political science journals," (*APSR, AJPS, BJPS*, and *JOP).* But while expanding the data's date range strengthens the claim that "taught IR" is different than "published IR," in this later case the update reduces the differences between IR and general interest journals. Since 2006 articles published in *IO, ISQ, WP*, and *IS* have become more similar to articles published in the general political science journals. Specifically, the IR journals have published a higher proportion of articles that are quantitative and positivist than in the pre-2006 era and have reduced the number of descriptive, formal, and analytic non-formal theory papers. So, by this measure, articles published in top IR journals look more like articles published in general interest journals, shrinking (but not eliminating) the purported gap between IR and political science.

Figure 6. Methods of Frequently-Taught Articles vs. All Published Articles

*Colgan's original table amended with new column for updated TRIP Journal Article Database

	2008-2013	1980-2006	*2007-2012
Method	**Freq Taught**	**% Published**	**% Published**
Quantitative	22%	34%	51%
Qualitative	34%	35%	32%
Formal	21%	13%	9%
Analytic & Non-Formal	38%	15%	10%
Experimental	0%	4%	3%
Descriptive	0%	10%	6%
Counterfactual	1%	1%	0%
Policy Analysis	0%	5%	2%

Colgan targets the "IR field seminar" in the top 65 U.S. PhD programs as the most relevant source for data on impact through teaching. This seems a good place to start, but we note several features of the sampling strategy that inhibit valid inferences and/or allow more noise than one might like around TIS estimates. Most obviously, the sample could be improved by collecting more than just one syllabus from one core IR course taught in one

semester at each university at some time over a seven-year period. The TIS sample is likely not representative of what any given department might include in its IR field seminar (much less its PhD curriculum) because the content of that syllabus almost certainly varies depending on the specific tastes of the instructor for that selected semester. For example, at UCSD in the early 1990s the "core seminar" in IR was actually two courses (the "system" course and the "unit" course), that were taught by some combination of John Ruggie, Lisa Martin, Peter Cowhey, David Lake, and Peter Gourevitch. Unsurprisingly, those were very different syllabi depending on the instructor!

Since the large variation observed in the olden days at UCSD did not seem like a unique situation, we emailed faculty members at two of the "outlier" departments (UVA and Northwestern) as measured by TIS 1.0. For the semester he analyzed, Colgan catalogued all the readings and concluded that "one of these universities, Northwestern, does not teach a single one of these canonical readings in its core IR course." But had Colgan happened to have sampled the syllabus from the very same course taught the following semester, he would have found that Karen Alter teaches 9 of the 10 most popular readings found on other IR syllabi, rather than 0 of 10, which he found on the syllabus taught in a previous semester. Similarly, while Colgan accurately reports that the UVA syllabus he analyzed "assigns almost nothing from *JCR* or *AJPS*," a second syllabus for the same course (co-taught by John Owen and Todd Sechser), but from a different year within the sampling frame, assigns seven different articles from either *JCR* or *AJPS*. Thus, the current measures provided by TIS are unnecessarily narrow and likely suffer from various types of measurement error. Colgan could improve the validity of TIS by including all IR field seminar syllabi used by a given department over any specified period of time.

Since TIS seeks to measure impact as represented by what is taught within the discipline, one could broaden the measure (or proliferate related measures) in a number of other ways, including, but not limited to: (1) measuring what readings are assigned in all PhD seminars rather than simply the field seminar; (2) expanding the date range to generate a larger sample; (3) analyzing what is taught in MA or BA courses to capture impact earlier and more broadly in the educational process; (4) increasing the number of PhD programs covered beyond the U.S. top 65, or beyond the U.S. Incidentally, when we asked all IR scholars how they organized their PhD field seminar in IR, their responses (254 of them) were broadly consistent with Colgan's findings -- instructors tended to organize the seminar around the big paradigmatic ideas that show up most frequently in Colgan's TIS measure. So, there is some additional indirect evidence that Colgan's results from the top PhD programs are representative of the broader population. While the sample could certainly be improved, Colgan's decision to start with top PhD programs makes good sense and helps to establish proof of concept.

The Challenge of Measuring Scholarly Impact: Triangulating What Counts

Whether citation counts *should* be used to evaluate scholarship is an ongoing debate, but compared to Colgan's TIS or reputational surveys, citation counts are likely going to be more widely applicable than what we currently have on hand. If one were sponsoring a competition for a lifetime achievement award in IR, or even a named professorship at a top 10 university, then a top 20 ranking on the TRIP survey or a top 20 ranking on a new TIS index might be modestly helpful. But for the rest of us who are trying to make decisions about whether to tenure someone who received his or her PhD a few years ago, TRIP and TIS are not all that useful, as both are extremely skewed toward the "top end" of the

distribution of scholars. Even if we dramatically expanded the number of syllabi, the number of courses covered, and the number of institutions covered, assistant professors will rarely appear. So, while TIS may be valuable for some things, including promotion to full professor at universities with very high standards, it is not all that useful as a substitute for citation metrics at tenure time, which is the decision point when such metrics likely matter most -- for good or ill.

While such metrics are unlikely ever to replace the subjective judgments of colleagues and external evaluation letters, they are systematically collected and over time we have learned more about the types of omissions and biases that are currently present in such measures. We know that citation counts do not tell the full story about the quality or impact of any piece of scholarship. This is why we can all benefit from the development of multiple different metrics, including Colgan's pioneering work on the "taught discipline." We encourage Colgan and his fellow travelers to continue improving TIS, since no single measure that we have today is sufficient to illuminate all the types of impact we might be interested in measuring and/or encouraging.

To Be, or Not to Be: 'Tis in Question

Cullen Hendrix
University of Denver

Jeff Colgan's Training Impact Score (TIS) is a new and interesting metric for determining "what counts" in international relations scholarship. The appeal is intuitive: as a scholarly community, we can judge what's important by looking at that which we see as vital for training the next generation of scholars. My take is that the TIS is a potentially useful tool, though I think there are some issues related to the sample from which it is derived and its comparison to Google Scholar, which I believe is still a better measure of total scholarly impact.[1]

Colgan's data "come from a systematic investigation of what we teach doctoral students of IR at U.S. universities." The data are culled from a convenience sample of 42 syllabi drawn from the top 65 graduate programs in Political Science, with some additional spadework done to boost representation of programs at the upper end of the rankings (Harvard, Duke, UC Berkeley). Crucially, Colgan only looks at "the core IR course for PhD students" because "it is the most comparable course across universities."

These choices are extremely consequential for a number of reasons. First, choosing to sample only political science PhD programs immediately reduces the scope of what constitutes international relations. The ISA is a big tent organization, with annual meetings attracting political scientists but also sociologists, economists, anthropologists and historians. This decision doesn't concern me personally, but I am quite sure that regular readers of *International Political Sociology*, many of whom self-identify as IR scholars and are ISA members, would find this choice disheartening.

Second, looking at the core course has merits and drawbacks. Appearing on a general IR seminar syllabus is certainly prestigious, but most scholarly contributions are not in the area of "general IR." Like any academic discipline, most scholarly contributions in IR occur several ramifications away from the trunk. Even articles appearing in *IO*, the most impactful journal identified in Colgan's analysis, are typically not "general", speaking rather to debates in international political economy, security, human rights, or global governance. Expanding the data analysis to include seminars on conflict, IPE, human rights, etc., would entail large investments in data collection but come closer to the goal of measuring "what we teach doctoral students of IR at US universities." Focusing on the core course probably overinflates the significance of canonical pieces that are important for "knowing the discipline" but that exert diminished or diminishing impact on the way current scholarship is conducted (Waltz' *Theory of International Politics* comes to mind). Colgan acknowledges this limitation and thus counsels caution in interpreting his results – yet still argues his measure is preferable to the emergent bean counter of record: Google Scholar.

Colgan launches a spirited criticism of Google Scholar, based on the asymmetry he finds between what is highly cited there and what is impactful according to his preferred measure

[1] This whole discussion is premised on the notion that what counts can be counted; I'm aware this view is not universally held, but since we're comparing different count-derived measures, I'll going to punt on this discussion.

and the TRIPS survey. Having spent some time looking at Google Scholar's data on IR scholars (Hendrix 2015), I found this critique the most provocative part of the article. Essentially, Colgan's argument is that many pieces that are highly cited according to Google Scholar do not rank highly on his measure of training impact, and this is taken to be evidence that Google Scholar is a flawed measure. My take-away is different: as a discipline, our perceptions of what work is influential are out of step with the way our work is used by broader academic and policy communities.

Why would some pieces be highly cited, according to Google Scholar, but not commonly taught? Setting aside the earlier point about the limitations of focusing on core IR seminars, Google Scholar casts a much wider net. Among Google Scholar's purported "evils" is that "Some citations come from academic scholars, but many do not: journalists, undergraduate students, policy analysts, and advocates also generate citations." To me, this is a feature, not a bug. I'm a little surprised that given an already narrow definition of the IR discipline that citations accrued in other disciplines were not added to the implicit "has no value" list. Google Scholar has noted holes – David Samuels found it systematically undercounts citations in books, (Samuels 2013) for instance (a drawback shared with SSCI) – but the nature of the bias is predictable, and it seems to err in the direction of inclusivity. If the price of catching citations to IR work in important amicus briefs, policy documents, and IGO reports is netting a few citations to undergraduate theses, I'm more than willing to pay. If we are truly concerned with bridging the gap between the ivory tower and the "real world" – as many of us are (here, here, and here, for instance) – adopting a measure of impact that excludes scholarly references from the latter risks marginalizing policy-relevant work in the former.

Google Scholar is then compared unfavorably (and the TIS favorably) to the TRIPS survey results on journal impact. That the TRIPS survey and the TIS are concordant doesn't surprise me: for most IR scholars, their IR survey course was the last time they discussed the broad field of IR in a PhD seminar. The vast majority of TRIPS respondents (75%, according to the 2012 *TRIP Around the World* report) do not teach PhD-level courses. I'd be willing to bet the majority of people teaching in IR PhD programs have not taught their institution's core IR seminar. This leads me to believe – or at least guess – that the two measures are similar largely because they are not independent of one another.

Google Scholar certainly doesn't need me as a champion; arguing against Google Scholar as the emergent norm for measuring scholarly impact (at least via citation metrics) is tilting at windmills. I'm deeply sympathetic to Colgan's point that knowing something about what contributions constitute the "core" of political science's take on IR matters, but I don't agree it matters more than a more open-ended measure of impact, or that general citation metrics are flawed because what appears influential according to them are not what appears influential on syllabi.

Here's my bottom line: the TIS is a nice measure of what is considered important, foundational, scope-defining stuff future IR scholars will need to know if they are to operate and converse with others in the political science wing of the international relations discipline. Thus it is useful, and I hope some of my papers eventually earn nice TIS scores. But it is an insular definition of "what counts." It accepts the premise that we are a hermetically sealed intellectual culture, and that our primary concern and therefore metric of impact, which affects decisions about hiring, tenure and promotion, should be based on influencing the training of the next generation of scholars. That's not a premise I accept, but it's a useful complement to other metrics of impact.

WHAT DO WE REALLY MEASURE WHEN WE TALK ABOUT "SCHOLARLY IMPACT"?

Jelena Subotic
Georgia State University

I read Jeff Colgan's article with great interest. At issue here is what do we mean when we talk about "scholarly impact." As Colgan demonstrates, there are many different ways to measure "impact," and the most common metric, publication citations, is deeply flawed. Colgan's analysis points to tremendous "noise in the system" in the most widely used citation metric, Google Scholar. After reading his analysis, I am beginning to think that using Google Scholar as measurement of scholarly impact is not much better than using the number of LinkedIn contacts a person has as measurement of career success. To correct for Google Scholar problems, Colgan introduces a new metric, which measures research impact by capturing the frequency with which scholarly articles are assigned in graduate IR syllabi.

I am certainly sympathetic to Colgan's project and find great value in exposing the distortions that the focus on citations creates. The research impact dynamic, however, operates within a much broader professional, sociological, and political environment of US academia, and there are some significant structural issues that influence perceived "impact," which Colgan's article leaves unproblematized.

The first issue has to do with using graduate syllabi as an authoritative source for research impact. As Colgan himself acknowledges, graduate syllabi themselves are ripe for critical treatment. They systematically overrepresent articles by male authors (Colgan 2015), articles published in U.S. journals, and as my own research shows, articles with a rationalist epistemological approach (Subotic 2017). The content of graduate syllabi can just as much be the result of an academic self-fulfilling prophecy, where only a certain kind of scholarship is taught in the top schools because only a certain kind of scholar is hired into these top schools (more of these findings are reported in Subotic 2017). There seems to be an assumption of meritocracy here (research included in the graduate syllabi is of top "quality") that is deeply problematic and ignores the profound structural inequalities that underpin the academic system such as, for example, its core-periphery structure (Clauset et al. 2015), or its oligarchic nature (Oprisko et al. 2013).

Further, the process of constructing a graduate syllabus, as Colgan also acknowledges, is prone to network effects, staleness (once prepared, syllabi may be updated on the margins, but the core structure typically remains the same), as well as similar elite-distorting effects discussed above. At most research institutions, especially the top-ranked ones that are the subject of this analysis, the incentives for publishing dwarf any incentives for quality teaching. This will lead time-crunched faculty to pay much less attention to syllabi research and assign already known works, or mimic syllabi from other peer institutions, and not spend the time needed to truly research new innovative work. It is hard to see how we can ignore these professional practices in syllabi analysis.

Finally, the focus on research "quality" and "impact" needs to take into consideration the broader professional, social, and political environment in which scholars work. This

obsession with numerical measurements such as Google Scholar follows the corporatization of universities, where an easily identifiable number can be shown to university "stakeholders," such as legislators or private donors, as measurement of scholarly "value." This practice has become so normalized that there now exists a "faculty productivity monitoring company" called *Academic Analytics*, which provides a proprietary index of faculty "productivity" composed of publications, citations, and grants, but neither teaching nor service (for a recent controversy at Rutgers University involving the use of *Academic Analytics*, see Flaherty 2015).

While Colgan is right to focus on finding a *better* metric, I would like us to reflect a bit deeper on what such metrics really tell us about the work we do, the perceived "value" our colleagues and society at large assign to our work, and the consequences of such instruments for the nature and integrity of our scholarship, and the professional environment in which we work.

Metrics Wars: How do we measure impact and quality in IR?

Jeff Colgan
Brown University

I am grateful to Jelena Subotic, Cullen Hendrix, Mike Tierney, and Katie Paulson-Smith for their lively and thoughtful engagement with my article, "Where is IR going?" All of them offer a mixture of intelligent critiques, encouraging words, and insightful extensions to my argument and empirical findings.

Subotic and Hendrix each critique the notion that syllabi from top U.S. universities are a good basis for measuring research influence, but they do so from opposite angles. Subotic points out that those syllabi probably give too little attention to heterodox theory (a concern shared by Kate McNamara 2009, Amitav Acharya & Barry Buzan 2007, and others), suggesting the method is not scholarly enough. Hendrix argues the approach is *too* scholarly, not giving sufficient attention to research used outside the ivory tower. Both arguments raise valid concerns about the method. Too much focus on mainstream scholarship will stifle innovation and new ideas. But fixing this problem isn't easy. There is a real danger in asking any one metric to do too much.

As I pointed out in my article, Google Scholar (GS) and the Training Influence Score (TIS) offer very different perspectives on research influence:

> Ninety-eight articles have more than 500 GS citations each but appear on fewer than three syllabi in the sample. Conversely, 111 articles appear on at least three syllabi but have fewer than 500 GS citations each. Again, a narrow focus on GS citations would privilege one set of articles at the expense of the other, when it is not obvious that it would be justified. Only 68 articles appear frequently on syllabi (at least 3 courses) and rank well on Google Scholar (at least 500 GS citations). In sum, Google Scholar provides a rather incomplete representation of research influence if viewed in isolation.

GS and TIS differ, in part, because they are not measuring a single underlying concept but instead some combination of what I label "impact" and "quality." Impact has to do with how many people are using or debating the research. Quality, on the other hand, includes accuracy, novelty, and breadth of applicability ("does this change the way I look at a lot of important things?").

All of the measures that we discuss in this symposium – citation-based metrics like GS, syllabi-based ones like TIS, or survey-based metrics like those produced by the Teaching and Research in International Policy (TRIP) project – measure some combination of quality and impact, though the exact balance between those ingredients might be different across metrics. My hunch is that GS is much better at measuring impact than it is quality.

Hendrix offers the most trenchant criticisms of my analysis, and a spirited defense of Google Scholar. My concern is that he conflates impact and quality. He agrees that we should have multiple measures but implicitly assumes (Hendrix 2015) that the measures are

all getting at a single thing called "productivity" or "influence." To me (and others), that is where the danger lies (Mugge 2015; Jackson and Nexon 2015).

This view of Google Scholar leads directly to what I call the "Thomas Friedman problem." Friedman has at least ten times as many GS citations as Hendrix and I do, combined. He has way more than Jim Fearon or Beth Simmons or any other IR scholar. What does that actually mean? Does Friedman have more "impact" than we do, in the sense of reach and popularity of his ideas? Absolutely. GS seems to capture this well. But does that mean Friedman is good at causal inference, or providing robust explanations of why things happen as they do in world politics? I'm much less sure of that.

What makes Google Scholar (or TIS or TRIP) valuable is the extent to which it disrupts the Old Boys Network (Saideman 2013) that, at least allegedly, used to characterize political science. If GS helps female and minority researchers, or scholars who graduated from lower-ranked universities, rise into the top echelons of the academy, so much the better. I'm not actually sure how much it does that, but it seems at least possible that metrics help counter some of the implicit biases that affect hiring, promotion, and tenure decisions. That said, let's not forget that Google Scholar introduces some serious biases of its own.

I can't neglect the fantastic response piece by Tierney and Paulson-Smith. They went above and beyond the call of duty. I especially like that they urge us to be careful about our inferences from a dataset with just one syllabus in one year from each school, a concern that I certainly share. I view my work as more of a "proof of concept" than the last word on what we can learn from syllabi. As Tierney and Paulson-Smith point out, one way to improve the product would be to increase the number of syllabi in the dataset.

One group of researchers has taken this approach to its logical extreme, collecting a million syllabi, as reported in the NY Times (Kalaganis and McLure 2016). That work is just being released, so I'm not clear on the details, but it looks like undergraduate syllabi are the majority in their sample. That approach has strengths and weaknesses, too: I suspect that, on average, there are more "provocative but low quality" readings assigned in undergrad classes than in core graduate seminars. That means that it might be less valuable as an indicator of research quality than one that draws only on graduate syllabi. Still, I suspect we will learn a lot from their project.

I'm indebted to Subotic, Hendrix, Tierney, and Paulson-Smith. All of us have substantive interests in world politics that take up the bulk of our time and attention. But spending a little time thinking about how we, as scholars, operate our own business turns out to be both fun and insightful. I hope the discussion will continue.

References

Acharya, Amitav, and Barry Buzan. 2007. "Why Is There No Non-Western International Relations Theory? An Introduction." *International Relations of the Asia-Pacific* 7 (3): 287–312. doi:10.1093/irap/lcm012.

Aggarwal, Vinod K. 2010. "I Don't Get No Respect:1 The Travails of IPE2." *International Studies Quarterly* 54 (3): 893–95. doi:10.1111/j.1468-2478.2010.00615.x.

Campbell, Peter, and Michael C. Desch. 2013. "Rank Irrelevance." *Foreign Affairs*, September 15. https://www.foreignaffairs.com/articles/united-states/2013-09-15/rank-irrelevance.

Catalini, Christian, Nicola Lacetera, and Alexander Oettl. 2015. "The Incidence and Role of Negative Citations in Science." *Proceedings of the National Academy of Sciences* 112 (45): 13823–26. doi:10.1073/pnas.1502280112.

Clauset, Aaron, Samuel Arbesman, and Daniel B. Larremore. 2015. "Systematic Inequality and Hierarchy in Faculty Hiring Networks." *Science Advances* 1 (1): e1400005. doi:10.1126/sciadv.1400005.

Cohen, Benjamin J. 2010. "Are IPE Journals Becoming Boring?" *International Studies Quarterly* 54 (3): 887–91. doi:10.1111/j.1468-2478.2010.00614.x.

Colgan, Jeff D. 2015. "New Evidence on Gender Bias in IR Syllabi." *Duck of Minerva*. http://duckofminerva.com/2015/08/new-evidence-on-gender-bias-in-ir-syllabi.html.

———. 2016. "Where Is International Relations Going? Evidence from Graduate Training." *International Studies Quarterly* 60 (3): 486–98. doi:10.1093/isq/sqv017.

Curry, Stephen. 2017. "Sick of Impact Factors." *Reciprocal Space*. Accessed August 1. http://occamstypewriter.org/scurry/2012/08/13/sick-of-impact-factors/.

Flaherty, Colleen. 2015. "Refusing to Be Evaluated by a Formula." https://www.insidehighered.com/news/2015/12/11/rutgers-professors-object-contract-academic-analytics.

Garand, James C., and Micheal W. Giles. 2003. "Journals in the Discipline: A Report on a New Survey of American Political Scientists." *PS: Political Science and Politics* 36 (2): 293–308.

Hagmann, Jonas, and Thomas J. Biersteker. 2014. "Beyond the Published Discipline: Toward a Critical Pedagogy of International Studies." *European Journal of International Relations* 20 (2): 291–315. doi:10.1177/1354066112449879.

Hendrix, Cullen. 2015. "Google Scholar Metrics and Scholarly Productivity in International Relations." *Duck of Minerva*. http://duckofminerva.com/2015/08/google-scholar-metrics-and-scholarly-productivity-in-international-relations.html.

Karaganis, Joe, and David McClure. 2016. "Opinion | What a Million Syllabuses Can Teach Us." *The New York Times*, January 22, sec. Opinion. https://www.nytimes.com/2016/01/24/opinion/sunday/what-a-million-syllabuses-can-teach-us.html.

Lake, David A. 2011. "Why 'isms' Are Evil: Theory, Epistemology, and Academic Sects as Impediments to Understanding and Progress." *International Studies Quarterly* 55 (2): 465–80. doi:10.1111/j.1468-2478.2011.00661.x.

Maliniak, Daniel, Ryan Powers, and Barbara F. Walter. 2013. "The Gender Citation Gap in International Relations." *International Organization* 67 (4): 889–922. doi:10.1017/S0020818313000209.

McNamara, Kathleen R. 2009. "Of Intellectual Monocultures and the Study of IPE." *Review of International Political Economy* 16 (1): 72–84. doi:10.1080/09692290802524117.

Mugge, Daniel. 2015. "The Collateral Damage of Performance Metrics." *Duck of Minerva*. http://duckofminerva.com/2015/08/the-collateral-damage-of-performance-metrics.html.

Nexon, Daniel H. 2011. Review of Review of Realist Constructivism: Rethinking International Relations Theory,; Rational Theory of International Politics, by J. Samuel Barkin and Charles L. Glaser. Perspectives on Politics 9 (4): 903–5.

———. 2015. "Special Event: 'The End of IR Theory' Symposium." *Duck of Minerva*. Accessed August 1. http://duckofminerva.com/2013/09/special-event-the-end-of-ir-theory-symposium.html.

Nexon, Daniel H., and Patrick Thaddeus Jackson. 2015. "Academia Isn't Baseball." *Duck of Minerva*. http://duckofminerva.com/2015/08/academia-isnt-baseball.html.

Oprisko, Robert L., Kristie Lynn Dobbs, and Joseph DiGrazia. 2013. "Pushing Up Ivies: Institutional Prestige and the Academic Caste System." *Georgetown Public Policy Review*. August 21. http://gppreview.com/2013/08/21/pushing-up-ivies-institutional-prestige-and-the-academic-caste-system/.

Østby, Gudrun, Håvard Strand, Ragnhild Nordås, and Nils Petter Gleditsch. 2013. "Gender Gap or Gender Bias in Peace Research? Publication Patterns and Citation Rates for Journal of Peace Research, 1983–2008." *International Studies Perspectives* 14 (4): 493–506. doi:10.1111/insp.12025.

Rathbun, Brian. 2017. "Stuff Political Scientists Like #9 — Being Liked, or Citations." *Duck of Minerva*. Accessed August 1. http://duckofminerva.com/2011/09/stuff-political-scientists-like-9-being_19.html.

Sabarantam, Meera. 2014. "Why Metrics Cannot Measure Research Quality: A Response to the HEFCE Consultation." *The Disorder Of Things*. June 16. https://thedisorderofthings.com/2014/06/16/why-metrics-cannot-measure-research-quality-a-response-to-the-hefce-consultation/.

Saideman, Steve. 2013. "Saideman's Semi-Spew: The Good Old Days in Academia." *Saideman's Semi-Spew*. December 11. http://saideman.blogspot.com/2013/12/the-good-old-days-in-academia.html.

Samuels, David. 2013. "Book Citations Count." *PS: Political Science & Politics* 46 (4): 785–90. doi:10.1017/S1049096513001054.

Sjoberg, Laura. 2015. "Why I Don't Give a Shit about My H-Index." *RelationsInternational*. August 17. http://relationsinternational.com/why-i-dont-give-a-shit-about-my-h-index/.

Subotic, Jelena. 2017. "Constructivism as Professional Practice in the US Academy." *PS: Political Science & Politics* 50 (1): 84–88. doi:10.1017/S1049096516002201.

Wemheuer-Vogelaar, Wiebke. 2013. "About 'The Gender Gap in IR and Political Science.'" *IR Blog*. August 31. http://irblog.eu/gender-gap-ir-political-science/.

Wight, Colin, Lene Hansen, Tim Dunne, John J. Mearsheimer, and Stephen M. Walt. 2013. "Leaving Theory behind: Why Simplistic Hypothesis Testing Is Bad for International Relations." *European Journal of International Relations* 19 (3): 427–57. doi: 10.1177/1354066113494320.