Instructions and Tips for Citations Searches in Publish or Perish and Web of Science

This document provides a modestly organized list of instructions, tips and correspondence among the research team for the Nosek, Graham, et al. (2010) investigation of citation impact (http://briannosek.com/papers/citations/). This may be helpful for other research teams that are thinking about replicating the approach and analysis in other disciplines. There was a good deal more correspondence than the items below. These capture what we perceive as the central issues in conducting effective searches with *Publish or Perish* software (as of Fall 2009).

The key issues are:

[1] defining the sample clearly before commencing the citation data collection

[2] gaining expertise in conducting good searches with the citation software – that includes how to find as many relevant citations as possible, how to spot errors, and how to clean citations (e.g., removing people with similar names).

[3] sharing that knowledge among the search team so that the same cleaning standards are used – i.e., avoid introducing systematic biases based on the person doing the search.

Initial instructions

Step 1: Identify qualifying members of Department (complete for all Departments before citation search)
1. Confirm that Department has a Social area
2. Gather names of core faculty - exclude secondary appointment and emerti
3. Insert core faculty names, institution, year of PhD, gender, and ethnicity in "People" worksheet
4. If there are ambiguities or unknowns follow up with someone in the Department (or individual him/herself)
5. Enter summary information about Department on "institutions" worksheet - your last name, Y for social area, # of core faculty

Common challenges
1. out of date website (new faculty not posted, retired or departed faculty not removed)
2. Department and personal websites do not list year of PhD; core faculty hard to distinguish from secondary appointments
3. Make sure "publishing name" is listed in spreadsheet (e.g., Anthony Greenwald, not Tony Greenwald)
4. Pay attention that person does not have multiple publishing names (usually noted on CV or obvious if large gap in pubs in years since PhD)

After all of these are completed by all collaborators, we will agree on a 1 week span for completing Step 2

Step 2: Citation count for each faculty member
1. Install Harzing's Publish or Perish (http://www.harzing.com/pop.htm)
2. Primary search is on person's first and middle publishing initial and last name, all in quotes: for example "TD Wilson" (including the quotes)
3. Remove all articles that do not belong to that author (challenging authors may need to introduce alternate strategies here)
4. Remove all articles with 0 citations
5. Save search results (naming convention: "TD.Wilson.Virginia.csv", and put all Virginia faculty searches into a folder called "Virginia")

Finding errors - missed articles, errant counts, double counts
1. Search on alternate names (e.g., first and last name in quotes, "Timothy Wilson" is main alternative - always use quotes)
2. Compare with Web of Science citation search to see if articles were missed, or any wildly discrepant counts on particular articles
3. Save alternates/final search if it generates a different result than the initial search (naming convention for saved file: TD.Wilson.Virginia.correction)
4. Log description of errors and corrections process as it deviates from saved search results in "Errors and corrections"
5. Enter final results in "People": total citations, years, h-index, e-index, hm-index (and Web of Science [WoS] total citation count if done as cross check)

Common challenges
1. Person may have published under two different names (e.g., if person changed name after getting married)
2. Person may have been inconsistent about using middle initial: "TD Wilson" will miss "T Wilson" publications, though "Timothy Wilson" might catch them
3. Person has a common name making it challenging to cull their publications: can restrict to subfields, exclude names, and use alternate naming e.g., "First Last" (careful)
   a. subfield restrictions - dropping physics and chem is usually safe, sometimes engineering and bio are safe, others are risky if person ever had more applied pubs
   b. "exclude names" can be quite helpful
   c. highest citation counts appear to occur with initials and last name as compared to first/last name, so watch them carefully of papers that only appear in one
   d. manually excluding articles with the standard search "TD Wilson" will be best strategy for most searches, main problems are common names
4. Publications not found when person's name is searched (e.g., N & Wilson, 1977 does not show up when searching "Timothy Wilson", but does with "TD Wilson")
5. Some erroneous results (e.g., Ed Diener search reveals a 1998 paper as highest cite (4454) that is actually more like 33 cites, be alert for outliers and document removal)

<u>Two instructional search examples (by Nosek)</u>

1. Here is an instructive case example of a complicated search and recalculation of the citation indicies.

Reuben M. Baron (try searching "RM Baron") is the first author of perhaps the most cited article in psychology's history - the mediator-moderator paper.

When you do the search you will notice that the moderator-mediator paper appears dozens of times in the search output. Its enormous number of cites leads to many typos in the cites. In fact, it is the 1st, 2nd, 4th, 6th, and 12th most cited of Baron's papers. This is likely the most complicated example of this (David Kenny will be similarly complicated).

To correct for this, we need only prevent the same paper from being counted multiple times in the "h" calculation. Remember, h is the highest value for which h is the number of papers that have been cited at least h times is still true. However, those typo cites should still be included in his overall citation count.

How to fix:

[1] First remove all non-Reuben Baron cites and all 0 cites (there is another RM Baron in biochemistry).

[2] Record the h and hm values after removing all extra mod-mediator cites (except the 1st) that have an impact on the h calculation. This results in h=20, hm=14.58).

[3] Note which article is the last one included in the h (Kelman & Baron - rank = 25) for later calculation.

[4] Next, bring the alternate versions of the mod-med paper back to get the total citation count (N = 18984). Record that in the spreadsheet.

[5] Finally, drop *all* papers past the last one that contributed to h (from step 3) and record the total citations count (N = 18316). (if you are fastidious - you will retain the alternate versions of the mod-med paper down the whole list (and other duplicates of papers in the "h-set", but this should be a minor influence.

[6] The number from 5 helps calculate a good e-index estimate [e-squared = 18316 - h-squared (20*20)]. so e = sqrt(18316-400). Record that value in e. Note in the comments when you needed to follow a procedure like this for recalculation. and provide the numbers that don't have a place in the columns, such as the value calculated in step 5.

2. Here is an example of a particularly challenging search and some steps that I took for narrowing this to be as efficient (and accurate) as possible.

Mark Snyder is productive and has had a long career. He doesn't use a middle initial when publishing and there are a lot of "snyders".

[1] Entered "M Snyder" as search - more than 1000 articles and lots of obvious stuff that wasn't his

[2] Restricted to publication years that are possible for him 1965 to 2009

[3] Removed some fields that were obviously non-overlapping. Started with chem, engineering, and physics. Still way too many cites, so removed medicine and biology too - still more than 1000 searches returned (true with all restrictions below too)

[4] skimmed list and pulled out names with other initials to remove: "cm snyder", "wm snyder", "mk snyder", "mj snyder", "jm snyder", "km snyder", "em snyder", "me snyder", "mr snyder", "ml snyder", "hm snyder", "kimm snyder", "am snyder", "mc snyder", "mh snyder", "sm snyder", "ym snyder", "nm snyder", "dm snyder", "ma snyder" , "mf snyder", "mr snyder", "rm snyder", "dm snyder", "mp snyder", "rm snyder" [would have been easier if I did even more]

[5] dropped 0 citations

[6] went cite by cite and started to see easy patterns for identifying ones to remove/keep -- article topics, co-author patterns, Did this through "2 citations" and then estimated the proportion of "1 citation" ones that would not be Mark based on the approximate proportion of 2 and 3 citation ones that were not his and removed that many by selecting a set and clicking "uncheck selection"

[7] cross check against WoS for big citations

This was pretty effective and reasonably quick despite the complexity.

Other tips exchanged between searchers during search phase (Step 2)

For Publish or Perish:
Try to be as inclusive as possible: even if the person almost always publishes with a middle initial, searching for "TD Wilson" will exclude some citations because some will not list the middle initial. Start with "T Wilson" and refine from there.

For very common names, it may not be possible to search for just the first initial and last name without reaching PoP's 1000-paper limit. One method that is more restricted but picks up almost all papers by the author is to search for them using the middle initial OR the first name (e.g., "TD Wilson OR Timothy Wilson").

When doing a search for someone who doesn't use their middle initial, like "N Miller", you can exclude "N* Miller" to get rid of NE Miller, NS miller, etc. Unfrotunately you can't also exclude "*N Miller" to get rid of all the SN Miller, FN Miller, etc. to drop multiple inititals use multiple asterisks, like excluding "N* Miller, N** Miller, and N*** Miller.

When you use the keyboard to travel up and down in the Results list, pressing the space bar toggles the check mark on and off on the selected line.

In some of the harder cases, leaving in the middle initial skips a bunch of citations, and leaving it out makes cleaning VERY difficult (even after the other reducing strategies). On trick that seems to work reasonably well in those cases is to do this in the search line: "BA Nosek" or "Brian Nosek" The latter catches (most) of the single initial cases. It doesn't solve when both initials are common or the first name/last name is super common, but it can help.

Also, I ran into a problem where a senior faculty member (Ph.D. = 1970)' first paper, which was cited > 1,000 times, didn't show up in the search results for KK Dion (her husband is KL Dion and publishes similar stuff, so I couldn't leave out the middle edition). I looked up the co-authors' information for that specific paper and tried an OR search query that would pull up the most population citation. I finally figured out that while a single author + OR multi-author won't work, reversing the order would: ("K Dion" "E Walster") OR "KK Dion"

For very difficult searches (i.e., two authors with similar names who do similar research, difficult to determine whether papers are or are not theirs), it may be necessary to consult the authors' CV online and do a check for each paper (this is quick for new faculty, not so quick for those with long vitas).

For Web of Science:
The supplementary "Web of Science" search is just supposed to be a back-up check that top cited articles are not missing, and that the citation counts are in the right ballpark for the main analysis. In general, Google Scholar appears to find 1.5 to 3 times as many citations as does WoS. If you find that WoS has similar or more cites than Google Scholar, it is a good indicator that something is amiss in one of the two searches.

Because WoS is just a back-up check, and not the full analysis, it should be as efficient as possible. With common names, I am finding that the easiest search strategy is to search "Nosek B*" (no quotes) and then use the "subject areas" refine search on the left. Click "more options/values" and select all of the categories that could be relevant - all ones mentioning psychology, and then related fields, especially with what you noticed the person studies in the main search (e.g., health related ones if it is a social-health psychologist). With that list I click "create citation report" and can quickly remove any lingering refs that are not correct with the check-boxes on the left. We just need a "good" estimate of the total citation count, and then skim the highly-cited to see if any obvious big papers were missed in the other search (also offers an opportunity to notice errors of inclusion on google scholar - though remember that WoS does not include books, chapters, or some journals).