Spring 5-8-2020

Google Scholar Versions: Errors and Implications

Daniel S. Dotson
Ohio State University - Main Campus, dotson.77@osu.edu

Google Scholar Versions: Errors and Implications

Daniel S. Dotson
The Ohio State University
155 South Oval Mall
180E Orton Hall
Columbus, OH 43210
dotson.77@osu.edu

Abstract

Google Scholar combines versions of what should be the same item into a single record with multiple versions listed and a common citation rate for all versions. However, these versions are not always the same document. A study on the citations of theses and dissertations found unusually high citation rates for some titles. On closer examination, these titles had versions that were other formats, sometimes with additional authors. A close examination of highly cited theses and dissertations revealed that nearly half of the titles were considered versions of other different formats, often much shorter and sometimes multi-authored journal articles.

Keywords

Google Scholar, versions, citation metrics, formats, dissertations, theses

Introduction

When examining citation rates of electronic theses and dissertations (ETDs) in Google Scholar for a previous study on citation rates, some very high citation counts were encountered. Upon closer examination, it was determined the citation counts were inflated due to Google Scholar counting ETDs as versions of other items in multiple cases. Primarily journal articles, but other formats were labeled as versions of the ETD if they had the same title. In addition to format differences of versions, some versions had multiple additional authors beyond the original ETD author.

Such occurrences raise multiple questions to be answered:

- How often is this occurring for ETDs?
- What formats are also being claimed as versions of ETDs? Are there sometimes multiple formats?
- Are “PDF sites” like ResearchGate and Academia.edu showing up as versions?
- What disciplines are the most affected by this versioning issue?

And some more philosophical questions outside the scope:

- Are some scholars using citation rates for “their works” that are inflated due to this false versioning? For example, if an ETD is cited, but is listed as a version of a journal article, are others using this citation as a quality indicator? This would be impossible to determine without interviewing the scholars involved.
- If there are errors with versioning with ETDs, how many other errors may be occurring with versioning and citations? Separate projects could explore this. PDF sites could be an interesting option.
- The citation count does not indicate which version was cited. Which version was actually cited? This could be explored, but the items examined had over 10,000 total citations, which would have been a very long-term project.
- Why did authors not consider using a unique title to make sure their works were more easily distinguishable? This particular question gets at the root cause of the false matching, but is unable to be explored without interviewing the authors.

While Google Scholar and its citation tracking are covered in the literature, the issue of dealing with Google Scholar’s versions is not as widely covered. Beel and Gipp (2010) covered the issue of free versions of articles sometimes being listed in Google Scholar as the default. Some of these are on so-called “spammer” sites. These authors questioned multiple practices by Google Scholar, including assigning papers improperly to the wrong journal, counting citations from less scholarly sources (like Power Point slides), and counting citations to items other than the official version. This last one gets at the tendency for Google Scholar to lump what they consider to be different versions of the same work together into one citation count rather than counting citations from any source into one master citation count.

Martin-Martin, Orduna-Malea, Harzing, and López-Cózar (2017) studied Google Scholar items for citation rates and examined items with high number of versions. They found a low but significant correlation between how many versions there were for an item and its position in search results. Their study's focus was more on the relationship between citations and position in results. They do bring up the idea that improper linking of different items as versions of the same document could be a factor in items' positions in search results. In other words, it is a known issue and it has ramifications. They do not explore this issue deeply, however.

Pitol and De Groot (2014) likewise examined Google Scholar citations. While they found that items with higher number of versions did not necessarily have higher citation rates, items that had free online versions did show higher citation rates. This related to versions that are freely found online, although the authors focused on the legitimate sites where content is legally available.

However, Google Scholar versions does have another problem. Bodlaender and van Kreveld (2015) explained the difficulty of merging versions (that are truly versions of the same work) into a single record for that item. This showcases that the versions of the same paper are sometimes separate when they should not be, while the issue being explored later in this article will show that items are sometimes improperly merged.

Methods

The previous study that focused on ETD use identified 66 ETDs that had seen citation rates of 50 or more. These ETDs are re-examined in this study of versioning. The following information was recorded:

- ETD information: Title, Author, Department, Year, Downloads (from original study), Degree.
- The previous citation count.
- The new citation count (usually higher, sometimes lower).
- Number of versions listed in Google Scholar.
- Whether only ETDs were listed in the versions (this could include ProQuest versions, database indexing of ETDs, PDFs of the ETDs on other sites, etc.
- Whether a version was a book, journal article, or conference paper – and if so, how many authors were on the item.
- Whether there were versions on PDF sites, such as Academia.edu, ResearchGate, or elibrary.ru. Author count was recorded.

Results

Differing Formats

The ETDs had a high percentage of items, 45.45%, which had versions listed in Google Scholar that were another format. A majority of these versions were journal articles and not the actual ETD. Four conference papers and one book were also listed as versions. The most cited title had versions from more than one format (the only such case), with a conference paper/presentation and a journal article version. Its citation count was the only one to hit four digits.

The data for these 66 titles are summarized in Table 1 and detailed information is in Table 3 (Appendix).

Results	#	%
All versions ETD	36	54.55%
Items not just ETDs	30	45.45%
Book	1	1.52%
Journal Article	26	39.39%
Conference Paper / Presentation	4	6.06%
On some PDF site	43	65.15%

Table 1: Summary of ETDs' Versions

Differing Author Counts

In terms of the number of authors of the non-ETD versions, the average was 2.33 with the most common number being 2. See Figure 1 for the number of items per author count.

Figure 1: Author count and Corresponding Items

In 22 (73.33%) of the cases of ETDs with versions that are not ETDs, one or more authors may be getting citation counts that are inflated as citations may have been to the ETD, but are attributed to the improper grouping of non-version versions.

Disciplinary Differences

To determine if there were disciplinary differences in how ETDs had versions that were not true versions assigned in Google Scholar, the 66 ETDs were assigned by the author to broad subject areas based on the departments. One item was assigned to the Graduate School, which was the department assignment given to digitized older ETDs ingested into the OhioLINK ETD Center repository. See Table 2. Refer to Table 3 (Appendix) for the original department. The color coding for format and subjects align with items in Table 3.

	Book	Conf Paper / Presentation	Journal Article	Multi[1]	Only ETD	Grand Total
Arts/Humanities	1 9.09%		3 27.27%		7 63.63%	11 16.67&
Graduate School					1 100%	1 1.52%
Sciences		1 4.17%	14 58.33%	1 4.17%	8 33.33%	24 36.36%
Social Sciences		2 6.67%	8 26.67%		20 66.67%	30 45.45%
Grand Total	1 1.52%	3 4.55%	25 37.88%	1 1.52%	36 54.55%	66

Table 2: ETDs by Discipline

[1]The single item that had both a journal article and conference paper as a version.

For both arts/humanities and the social sciences, the majority of ETDs from the items examined had versions that were only ETDs. The sciences, on the other hand, had a majority of items that had versions that were actually other formats, the majority being journal articles. The single Graduate School item was only an ETD. Arts/Humanities had the only item that had a version that was a book format.

Discussion

The differences between the ETDs and the other formats improperly listed as versions were the primary focus of this study. Enough items were found in this group to indicate Google Scholar is not always properly versioning items. A thesis or dissertations with a single author and many pages is clearly not the same thing as a journal article or conference paper with multiple authors and fewer pages.

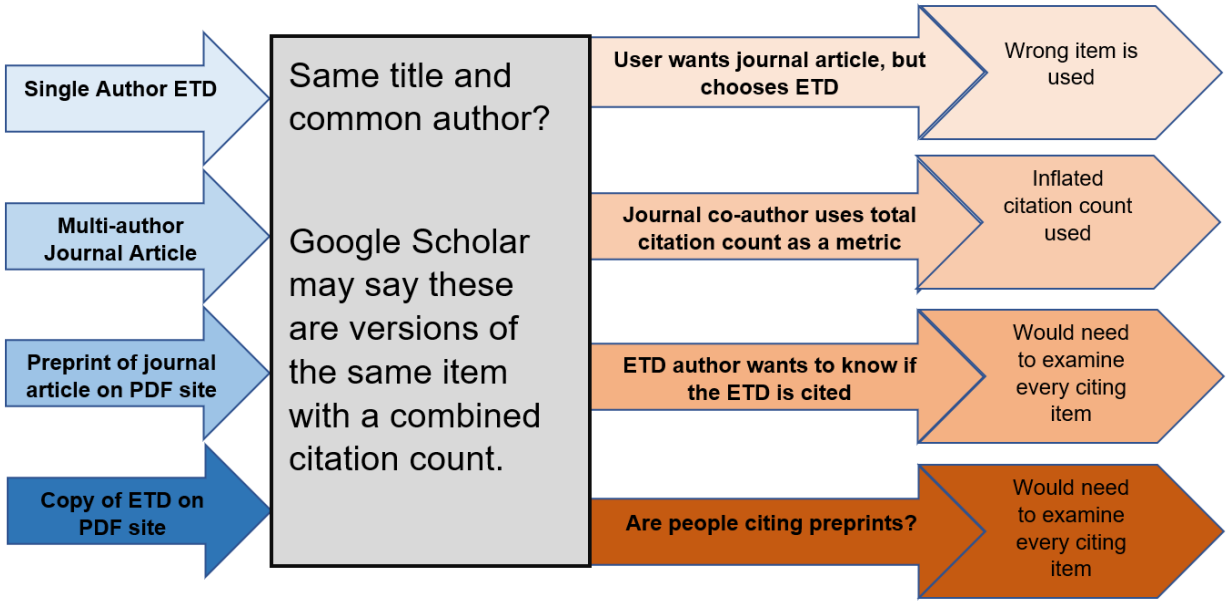

Figure 2: The Problems of Improper Versioning

Figure 2 illustrates the implications that can occur due to improper versioning. This leads to many potential issues. In addition to the obvious that searchers may select the wrong item and use it instead of the intended one, there is also a question of citation metrics. Google Scholar combines citations to all versions in one count. That count includes items that are potentially not the work of some authors in such cases. While it seems likely that a majority, most, or even all of the citations are to the non-ETD formats, it remains possible that some items in the citation count are to the ETD. In such cases, authors listed on the other items may be counting these numbers to a citation that in reality is to the ETD author only. The more extreme example where more than two formats (ETD, journal article, and conference paper) are versioned by Google Scholar makes the issue even more complicated.

PDF sites were perhaps one of the more interesting aspects of the versioning issue. Some contained the ETD, while others contained the other format(s) listed among the versions. A total of 43 items were on such PDF sites. Seven of the titles have items on PDF sites were differing author counts, indicating that at least some of them were ETDs and some were the other format. This makes the versioning even more complicated.

The most common true versions of the ETDs in question were linked to the ProQuest Dissertations & Theses platform, where Ohio State's doctoral dissertations, but not masters theses, are submitted. This information was not tracked.

An extreme example that does well to illustrate the problem with version is an item that was an ETD versioned alongside a journal article, a conference paper, PDFs on various PDF sites, and also some citation-only listings. This is illustrated in Figure 3.

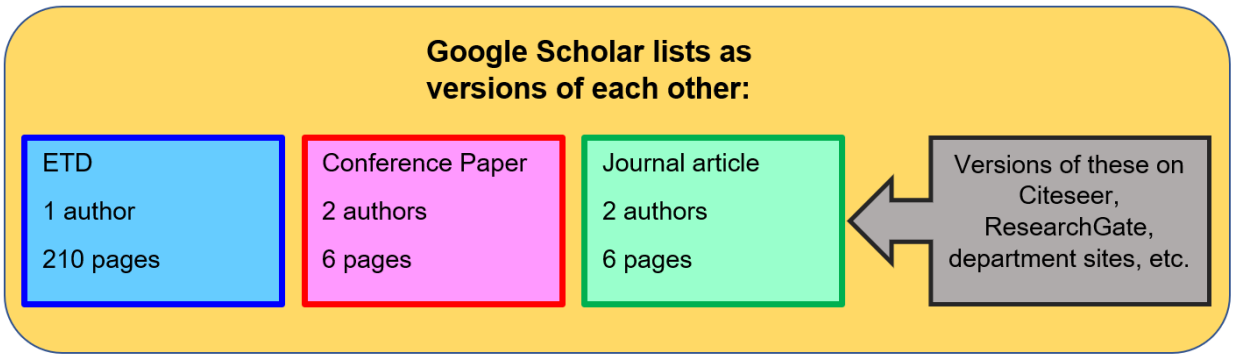

Figure 3: Many Versions not Versions

Conclusions

Based upon the findings, citation rates for items with multiple versions should be suspect. Nearly half of these highly-cited ETDs had versions that were other formats, leaving the question as to which item was actually cited. Due to titles being the same for the different formats and having a common author, these versions were intermingled, and their citation rates combined. This leaves one unsure, without examining every citation, which format got the citation. In some cases, multiple other authors may be getting citation counts in Google Scholar to which they are not entitled. Are people citing the ETD or the journal article? Are people citing the PDF on Academia.edu that is a PDF of the ETD or a PDF on elibrary.ru that is a PDF of the journal article? This level of complexity points to the inexactness of the citation counts. Authors should examine their citations more carefully in such cases where these versions are combined, sometimes improperly, rather than trust the citation count outright.

Finally, a question not originally explored was the tendency for versions that were not full text. A number of versions were simply citations to one or more formats. Some were databases that only indexed an item which was often the ETD. These versions were really not useful beyond the fact that the item exists and did not create a true version.

References

Beel, J., & Gipp, B. (2010). Academic search engine spam and Google Scholar's resilience against it. *Journal of Electronic Publishing, 13*(3). http://dx.doi.org/10.3998/3336451.0013.305

Bodlaender, H. L., & van Kreveld, M. (2015). Google scholar makes it hard – the complexity of organizing one's publications. *Information Processing Letters, 115*(12), 965-968. http://dx.doi.org/10.1016/j.ipl.2015.07.003

Martin-Martin, A., Orduna-Malea, E., Harzing, A., & López-Cózar, E.D. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics, 11*(1), 152-163. http://dx.doi.org/10.1016/j.joi.2016.11.008

Pitol, S.P., & De Groote, S.L. (2014). Google Scholar versions: Do more versions of an article mean greater impact? *Library Hi Tech, 32*(4), 594-611. https://doi.org/10.1108/LHT-05-2014-0039

Appendix

Title	Author	Dept[1]	Year	Downloads	Degree	Original Cite Count	New Cite Count	Versions	Formats beyond ETD	# authors	PDF sites	PDF site authors
Stability Analysis of Swarms	Veysel Gazi	Electrical and Computer Engineering	2002	5102	Doctoral	1076	1117	20	Journal article Conf Paper /Presentation	2	X	mixed
Short Text Classification in Twitter to Improve Information Filtering	Bharath Sriram	Computer Science and Engineering	2010	23540	Masters	786	865	11	Conf Paper /Presentation	5	X	mixed
The Phonetics and Phonology of Korean Prosody	Sun-Ah Jun	Linguistics	1993	3199	Doctoral	762	915	10	Book	1	X	1
The effects of the classroom flip on the learning environment: a comparison of learning activity in a traditional classroom and a flip classroom that used an intelligent tutoring system	Jeremy Strayer	Educational Studies	2007	34101	Doctoral	421	504	5	Only ETD	1		
Root resorption associated with orthodontic tooth movement: A Systematic Review	Belinda Jessica Weltman	Dentistry	2009	4679	Masters	375	415	12	Journal article	5		
A comparative analysis of energy management strategies for hybrid electric vehicles	Lorenzo Serrao	Mechanical and Aerospace Engineering	2009	8652	Doctoral	359	411	17	Journal article	3	X	3
A theoretical and experimental investigation of modulation sidebands of planetary gear sets	Murat Inalpolat	Mechanical and Aerospace Engineering	2009	6489	Doctoral	258	293	7	Journal article	2	X	mixed
Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control	Brandon D. Stewart	Psychology	2007	2004	Doctoral	214	232	16	Journal article	2	X	mixed
Forming of tailor-welded blanks	Frederick I. Saunders	Materials Science and Engineering	1994	3875	Doctoral	203	197	6	Journal article	2		
Toward the design of a computer-based interactive fantasy system	Brenda Kay Laurel	Theatre	1986	2284	Doctoral	200	206	2	Only ETD	1		
The effects of humor in persuasion	Dorothy Markiewicz	Psychology	1972	3941	Doctoral	195	205	6	Journal article	1	X	1
The five-factor model and career self-efficacy: general and domain-specific relationships	Robert Owen Hartman	Psychology	2006	13056	Doctoral	184	205	9	Journal article	2	X	mixed
Alienation and political apathy	Dwight Gantz Dean	Sociology	1956	2039	Doctoral	147	157	6	Journal article	1		
A uniform pressure electromagnetic actuator for forming flat sheets	Manish Kamal	Materials Science and Engineering	2005	4680	Doctoral	127	147	17	Journal article	2	X	1
Social capital and political consumerism: a multilevel analysis	Lisa Anne Neilson	Sociology	2006	3391	Masters	120	136	8	Journal article	2	X	mixed
Mechanics and mechanisms of ultrasonic metal welding	Edgar de Vries	Materials Science and Engineering	2004	13538	Doctoral	119	128	5	Only ETD	1		
A model of school success: instructional leadership, academic press, and student achievement	Jana Michelle Alig-Mielcarek	Educational Studies	2003	43092	Doctoral	117	79	5	Only ETD	1	X	1
Mechanisms of corrosion inhibition of AA2024-T3 by vanadates	Mariano Iannuzzi	Materials Science and Engineering	2006	3768	Doctoral	117	126	12	Journal article	2	X	1

Title	Author	Dept[1]	Year	Downloads	Degree	Original Cite Count	New Cite Count	Versions	Formats beyond ETD	# authors	PDF sites	PDF site authors
Dynamic melodic expectancy	Bret J. Aarden	Music	2003	2655	Doctoral	117	119	2	Only ETD	1		
Omega-3 fatty acids effect on wound healing	Jodi C. McDaniel	Nursing	2007	2796	Doctoral	116	126	17	Journal article	4	X	mixed
The conceptual structure of emotional experience in Chinese /	Brian King	Graduate School	1989	5282	Doctoral	111	118	4	Only ETD	1		
Estimation of the standard error and confidence interval of the indirect effect in multiple mediator models	Nancy Elizabeth Briggs	Psychology	2006	3412	Doctoral	110	115	5	Only ETD	1		
The emergence of distinctive features	Jeff Mielke	Linguistics	2004	2470	Doctoral	106	112	8	Only ETD	1		
A Lockean Theory of Intellectual Property	Adam D. Moore	Philosophy	1997	3415	Doctoral	105	114	6	Journal article	1		
A criterion-related validity test of selected indicators of musical sophistication using expert ratings	Joy E Ollen	Music	2006	4857	Doctoral	103	110	6	Only ETD	1	X	1
Relationships of selected factors and the level of computer use for instructional purposes by technology education teachers in Ohio public schools: a statewide survey	Mohammed Ibrahim Isleem	Teaching and Learning	2003	4166	Doctoral	102	111	6	Only ETD			
Development of a generalized mechanical efficiency prediction methodology for gear pairs	Hai Xu	Mechanical and Aerospace Engineering	2005	8340	Doctoral	99	111	4	Only ETD	1	X	1
Employee silence: Investigation of dimensionality, development of measures, and examination of related factors	Chad Thomas Brinsfield	Management and Human Resources	2009	8353	Doctoral	98	108	4	Only ETD	1		
Analysis of crack propagation in asphalt concrete using a cohesive crack model	Jia-Der Perng	Civil Environmental, and Geodetic Engineering	1989	3424	Masters	93	95	3	Journal article	2		
Student Attitude Toward STEM: Development of an Instrument for High School STEM-Based Programs	Mark Patrick Mahoney	Teaching and Learning	2009	3503	Doctoral	92	110	14	Journal article	1	X	1
An exploration of the factors associated with the attitudes of high school EFL teachers in Syria toward information and communication technology	Abdulkafi Albirini	Educational Studies	2004	4097	Doctoral	89	102	7	Only ETD	1	X	1
Chinese morphology and its interface with syntax	Xiang-ling Dai	Linguistics	1992	3202	Doctoral	88	93	4	Only ETD	1	X	1
Turkish college students' willingness to communicate in English as a foreign language	Yesim Bektas Cetinkaya	Teaching and Learning	2005	8824	Doctoral	86	102	7	Only ETD	1	X	1
Modeling and Control of a Hybrid-Electric Vehicle for Drivability and Fuel Economy Improvements	Kerem Koprubasi	Mechanical and Aerospace Engineering	2008	13501	Doctoral	85	92	7	Only ETD	1	X	1
Plasticity-Based Distortion Analysis for Fillet Welded Thin Plate T-Joints	Gonghyun Jung	Materials Science and Engineering	2003	5872	Doctoral	83	79	7	Journal article	2	X	1

Title	Author	Dept[1]	Year	Downloads	Degree	Original Cite Count	New Cite Count	Versions	Formats beyond ETD	# authors	PDF sites	PDF site authors
The Slaying of Lady Mondegreen, being a Study of French Tonal Association and Alignment and their Role in Speech Segmentation	Pauline Susan Welby	Linguistics	2003	2205	Doctoral	82	84	4	Only ETD	1		
Video compression and rate control methods based on the wavelet transform	Eric J Balster	Electrical and Computer Engineering	2004	2583	Doctoral	77	83	5	Only ETD	1	X	1
High school student's motivation to engage in conceptual change-learning in science	Lily Barlia	Teaching and Learning	1999	3390	Doctoral	76	84	5	Conf Paper /Presentation	2	X	1
Microphone based on Polyvinylidene Fluoride (PVDF) micro-pillars and patterned electrodes	Jian Xu	Mechanical and Aerospace Engineering	2010	13374	Doctoral	74	79	6	Journal article	4		
Religious democrats: democratic culture and Muslim political participation in post-Suharto Indonesia	Saiful Mujani	Political Science	2004	5147	Doctoral	73	71	5	Only ETD	1	X	1
Smile Esthetics from Patients’ Perspective for Faces of Varying Attractiveness	Chang Alexandra Chan	Dentistry	2011	2062	Masters	72	91	8	Journal article	7		
Job satisfaction of vocational teachers in Puerto Rico	David Padilla-Velez	Agricultural and Extension Education	1993	2034	Doctoral	68	66	5	Only ETD	1	X	1
Becoming a sports fan: understanding cognitive development and socialization in the development of fan loyalty	Jeffrey D. James	Educational Studies	1997	6650	Doctoral	67	65	4	Only ETD	1	X	1
How They Decide: A case study examining the decision making process for keeping or cutting music education in a K-12 public school district	Marci L. Major	Music	2010	4533	Doctoral	66	81	3	Journal article	1		
Describing and measuring the athletic identity construct: Scale development and validation	Thomas J Cieslak	Human Sciences	2004	18485	Doctoral	65	86	3	Only ETD			
Cylindrical FDTD analysis of LWD tools through anisotropic dipping layered earth media	Hwa Ok Lee	Electrical and Computer Engineering	2005	2971	Masters	65	69	11	Journal article	2	X	2
The Limits to Judicialization: Legislative Politics and Constitutional Review in the Iberian Democracies	Pedro C Magalhaes	Political Science	2003	3781	Doctoral	64	76	4	Only ETD	1	X	1
Tool Degradation Characterization in the Friction Stir Welding of Hard Metals	Brian Thomas Thompson	Materials Science and Engineering	2010	2052	Masters	62	70	6	Journal article	2		
Using high-probability request sequences to increase social interactions in young children with autism	Sunhwa Jung	Human Sciences	2003	6579	Doctoral	61	76	9	Journal article	3		
No child left behind: determining the impact of policy on music education	Kevin W Gerrity	Music	2007	3159	Doctoral	61	66	4	Journal article	1		
Supply Chain Resilience: Development of a Conceptual Framework, an Assessment Tool and an Implementation Process	Timothy J. Pettit	Business Administration	2008	2766	Doctoral	61	77	9	Only ETD	1	X	1
Fan loyalty : the structure and stability of an individual's loyalty toward an athletic team	Daniel Carl Funk	Human Sciences	1998	4980	Doctoral	60	60	4	Only ETD	1	X	1

Title	Author	Dept[1]	Year	Downloads	Degree	Original Cite Count	New Cite Count	Versions	Formats beyond ETD	# authors	PDF sites	PDF site authors
Joining enabled by high velocity deformation	Peihui Zhang	Materials Science and Engineering	2003	3988	Doctoral	60	60	3	Only ETD	1	X	1
The use of the internet among EFL teachers at the Colleges of Technology in Saudi Arabia	Ali M. Al-Asmari	Teaching and Learning	2005	4904	Doctoral	59	59	6	Only ETD	1	X	1
Exploring willingness to communicate (WTC) in English among Korean EFL (English as a foreign language) students in Korea: WTC as a predictor of success in second language acquisition	Seung Jung Kim	Teaching and Learning	2004	5380	Doctoral	58	71	5	Only ETD	1	X	1
Effects of self monitoring on the on-task behavior and written language performance of elementary students with learning disabilities	Laura Harkness Wolfe	Human Sciences	1997	2433	Masters	58	65	11	Journal article	3	X	3
A phonological study of some English loan words in Japanese	Mieko Ohso	Linguistics	1971	3546	Masters	55	56	2	Only ETD	1		
Electrochemical quartz crystal microbalance study of corrosion of phases in AA2024-T3	Younghoon Baek	Materials Science and Engineering	2002	2050	Masters	55	64	7	Journal article	2	X	2
The response of primary children to picture books	Barbara Zulandt Kiefer	Educational Studies	1982	2740	Doctoral	54	41	3	Only ETD	1	X	1
The role of enabling bureaucracy and academic optimism in academic achievement growth	Leigh McGuigan	Educational Studies	2005	2606	Doctoral	54	54	4	Only ETD	1	X	1
Modeling and control of a hybrid electric drivetrain for optimum fuel economy, performance and driveability	Xi Wei	Mechanical and Aerospace Engineering	2004	14269	Doctoral	52	50	6	Only ETD	1	X	1
Exploring organizational learning culture, job satisfaction, motivation to learn, organizational commitment, and internal service quality in a sport organization	Di Xie	Human Sciences	2005	9970	Doctoral	52	55	3	Only ETD	1	X	1
Factors affecting attitudes toward seeking and using rormal mental health and psychological services among Arab-Muslims population	Nasser Aloud	Social Work	2004	17190	Doctoral	50	54	7	Only ETD	1	X	1
Perceptions of service quality, satisfaction and the intent to return among tourists attending a sporting event	David J. Shonk	Human Sciences	2006	11400	Doctoral	50	61	5	Conf Paper /Presentation	1	X	1
Determinants of Shadow Education: A Cross-National Analysis	Darby E. Southgate	Sociology	2009	7887	Doctoral	50	60	7	Only ETD	1	X	1
On improving the accuracy and reliability of GPS/INS-based direct sensor georeferencing	Yudan Yi	Earth Sciences	2007	5559	Doctoral	50	48	5	Only ETD	1	X	1

Table 3: ETDs' Versions for Highly Cited ETDs

[1]The departments are further explored in Table 2, using this color coding for broader subject disciplines.