Bibliography profiling of undergraduate theses in a professional psychology program*

Perfil bibliográfico de las tesis de grado en un programa de psicología profesional

JAIME R. ROBLES,** EUGENIA CSOBAN-MIRKA*** and CRISTINA VARGAS-IRWIN****

Abstract

The bibliographic profile of 125 undergraduate (licentiate) theses was analyzed, describing absolute quantities of several bibliometric variables, as well as within-document indexes and average lags of the references. The results show a consistent pattern across the years in the 6 cohorts included in the sample (2001-2007), with variations, which fall within the robust confidence intervals for the global central tendency. The median number of references per document was 52 (99% CI 47-55); the median percentage of journal articles cited was 55%, with a median age for journal references of 9 years. Other highlights of the bibliographic profile were the use of foreign language references (median 61%), and low reliance on open web documents (median 2%). A cluster analysis of the bibliometric indexes resulted in a typology of 2 main profiles, almost evenly distributed, one of them with the makeup of a natural science bibliographic profile and the second within the style of the humanities. In general, the number of references, proportion of papers, and age of the references are close to PhD dissertations and Master theses, setting a rather high standard for undergraduate theses.

Keywords: licentiate thesis, undergraduate thesis, bibliographic profile, bibliometric indexes, robust confidence intervals, bootstrap, cluster analysis.

Resumen

Se analizó el perfil bibliográfico de 125 tesis de grado (licenciatura), mediante la descripción de cantidades absolutas de diversas variables bibliométricas, así como los índices de los documentos y el promedio de rezago de las referencias. Los resultados muestran un patrón consistente a lo largo de los años en las seis cohortes incluidas en la muestra (2001-2007), con variaciones, las cuales caen dentro de los intervalos de confianza robustos para la tendencia global central. El número medio de referencias por documento fue 52 (47-55 IC 99%); el porcentaje medio de artículos de revistas citados fue de 55% con una edad media de referencias de revistas de 9 años. Otros aspectos destacados del perfil bibliográfico fueron el uso de referencias en lengua extranjera (61% de media) y baja dependencia de los documentos abiertos en la *web* (2% de media). Un análisis de conglomerados de los índices bibliométricos resultó en una tipología de los dos principales perfiles, distribuidos casi por igual, uno de ellos con la marca del perfil bibliográfico de una ciencia natural, y el segundo dentro del estilo de las humanidades. En general, el número de referencias, la proporción de artículos y la edad de referencias están cerca de las disertaciones doctorales y las tesis de maestrías, estableciendo un estándar bastante alto para las tesis de grado.

Palabras clave: tesis de licenciatura, tesis de grado, perfil bibliográfico, índices bibliométricos, intervalos de confianza robustos, análisis "*bootstrap*", análisis de conglomerados.

* Corresponding author: Cristina Vargas-Irwin, Fundación Universitaria Konrad Lorenz, Carrera 9 Bis # 60 - 43, Bogotá, Colombia. Telephone: 3472311, Ext. 111. E-mail: cvargas@fukl.edu

** Universidad Católica Andrés Bello, Caracas, Venezuela.

*** Universidad Católica Andrés Bello, Caracas, Venezuela.

**** Fundación Universitaria Konrad Lorenz, Bogotá, Colombia.

Introduction

Undergraduate theses or dissertations are often considered the first complete exercise in scientific writing for psychology students. The process of writing a thesis is often studied from the perspective of the task-time dynamics or the personal traits associated with task completion (Klassen, Krawchuk, & Rajani, 2008; Klassen & Kuzucu, 2009; Rosario, Costa, Núñez, González-Pienda, Solano, & Valle, 2009; Seo, 2008; van der Hulst & Jansen, 2002). There is, however, another aspect of the undergraduate thesis process, which is the characterization of the resulting documents. Regardless of a variety of guidelines and standards for their elaboration, the thesis as a document has a set of empirical quantitative properties which may shed light on the types of references most frequently used in the research process, the most consulted journals in a given field, and the obsolescence rate of journals (Buchanan & Herubel, 1994; Vallmitjana & Sabate, 2008). Additionally, quantitative analysis of bibliographies, which is commonly referred to as bibliography profiling (Buchanan & Herubel, 1994), may provide feedback to existing guidelines for thesis elaboration or aid in the creation of new ones.

The bibliographic information contained in the citations constitutes one of the main aspects of the quantitative properties of the thesis as a document. The main aim of this paper is to describe the distribution of several bibliometric indicators for undergraduate theses, providing reference statistics useful for establishing standards for this type of document.

Studying undergraduate theses poses a series of challenges. One of them is the lack of uniformity of undergraduate programs, which vary widely across countries and institutions. Key differences are the length, course load and end-title awarded at the completion of the program. In this case, the theses studied are performed within the context of a five-year, full course load professional psychology program, resulting in a professional degree (licentiate), which allows the titleholder to start a lawful professional practice, or professional activities in general. This type of program is typical of most Latin American countries (Ardila, 1986). Even within this context, the undergraduate thesis is not an ubiquitous requirement. Most 3 or 4 year undergraduate programs do not result in a professional degree, and in most cases, when the thesis is performed, it is not a requirement but part of an optional honor program.

All these issues increase the difficulty of finding previous comparable studies or making generalizations based upon any of the existing studies, given the lack of uniformity of the programs and documents, and the scarce nature of the data.

Most bibliometric analyses of thesis have specific features and purposes, usually driven by intra-institutional aims. In many cases, the main purpose is to provide feedback to the originating institution as for the use of library, labs and other resources (Leiding, 2005; Walters, 2008).

An additional factor contributing to the difficulty of using a standard approach for the bibliometric study of undergraduate theses, is the fact that they are seldom published in major journals, indexed or cataloged in standard widely-available databases, thus allowing for citation analysis.

Other studies include undergraduate theses as part of wider bibliometric analysis, which encompass Master's and Doctoral Dissertations as well. It is clear that the academic process within a professional program (undergraduate) is different from a graduate program, especially in contexts in which the post-graudate degree *is not required* for professional practice. In this sense, the study of undergraduate thesis has to be specific and the characterization of these documents is useful to shape guidelines only if these studies focus on a particular kind of academic program.

Given those factors and contextual conditions, the bibliometric study of undergraduate thesis has to focus on within-document citation properties, instead of cross-citation analysis. Establishing the statistical properties of the distribution of the bibliographic entries in the thesis through bibliometric indexes can play a major role in shaping educational policy regarding theses standards.

Bibliometric indexes may also provide valuable insights into the diffusion and construction of knowledge within a specific field. Hargens (2000)

has proposed a classification of scientific disciplines based upon the relative importance given to foundational and current scholarship. On one end of the continuum lie those disciplines that emphasize current research and rarely acknowledge original sources, since they are usually fully assimilated into standard scientific practices. References in this type of field usually become obsolete more rapidly and new findings are promptly incorporated into subsequent work. On the other end of the continuum lie those disciplines that emphasize the interpretation of canonical work where the main task of the scholar is not the generation of new empirical findings, but rather the production of current insights into classical problems. This type of discipline exhibits extensive use of "orienting reference lists", that is, of sets of citations to well known classics aimed at supplying the reader with a conceptual framework which justifies the authors' particular approach to a problem, rather than on providing empirical evidence on the subject matter. While the former type of scholarship typically characterizes research in the natural sciences, the latter is more typical of the humanities.

Other characteristics usually accompany both ends of the typology: scholarly work in the humanities usually exhibits a more regional orientation, while research in the natural sciences tends to follow international standards and interests (Adams, 1998). In the same token, the main type of references used in the natural sciences are citations to research articles, while more frequent citing of books is characteristic of the humanities (Nederhof, Zwaan, Debruin, & Dekker, 1989). Finally, single-author scholarship is far more frequent in the humanities, where more emphasize is given to the production of literature aimed at the general public (Nederhof, 2005).

In view of the above, the main objective of the present study was to characterize the prevailing scholarship style of the undergraduate thesis in psychology for a particular Latin American undergraduate program. The group of documents studied in this paper are a result of a thesis supervisory system implemented in the school of psychology at Universidad Católica Andrés Bello (UCAB), in Caracas, Venezuela. This system required students in the last year of the undergraduate program to perform systematic assessments and reports of the completion of their thesis during that academic year. One of the required reports included bibliometric information, specifically that concerning the bibliographic references cited in the document. The indexes which can be derived from such information are thus within-document citation indexes.

Most standard cross-citation indexes have applications in evaluation of scientific output or even predictive value on the scientific production (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Hirsch, 2007; Mathur & Sharma, 2009; Thompson, Callen, & Nahata, 2009). Within-document indexes, on the other hand, focus in the description of the properties of the document, and not in the output or impact produced by the authors.

Those indexes fall across two main axes: year of citation and source differentiation. Most time-oriented indexed are related to shelf-life and half-life indicators as cross-citation indexes. In the present analysis, within-document time indicators must be used, focusing on average publication year of the cited references as the basis to build several time-related indexes.

Regarding source differentiation, various ratios for an undergraduate thesis were included: book/papers citation ratio, and topic diversity, which can indicate the orientation and depth of the reference search.

Method

Cohort description

Bibliographic reports archived by the thesis advising commission at UCAB were used to build the database used in this study. The reports range from 2001 to 2007, with a total of 125 theses, with the exception of 2005, due to unavailability reports for this cohort. The database was therefore built using data from 6 cohorts.

Data

In bimonthly reports, signed by the Advising Professor, the thesis authors reported the following

indicators: Total Number of References, Average Year of the References, Number of References to Journals, Books, Psychology References and Inter-Disciplinary References. The data from the final report was taken to be an accurate estimate of the bibliometric quantities in the final document. These reports were completed several weeks before the final document is presented. However, a previous study (Hargens, 2000; Robles, Csoban-Mirka & Vargas-Irwin, 2009) indicated that by the time of this report, a minimum of 80% of the references has been completed, and 75% of the thesis has 95% or more of their references completed. This means the absolute quantities might be slightly underestimated, however, the ratios and indexes can be considered representatives, as they are less dependent on those small changes in the absolute quantities.

Several quality assurance operations are performed on the dataset, include range checking, random verification with the final document and consistency analysis on previous reports. A total of 49 out of the 125 thesis studied (39.2%) were verified against the final document, finding a 92.5% consistency rate between the reports and the documents. At least one document from each cohort was verified.

Index definition and estimation

Several bibliometric indexes were computed from the data gathered by the Thesis Supervisory Board. These indexes may be divided into two main categories: source-oriented and time oriented indexes.

Source-based indexes

a. Source ratios: percentage of the total references compromised by each type of source. These percentages include: percentage of journal articles, of books, of citations of theses and of citations to open web documents.
b. Index of Cross-Disciplinary References, defined as the percentage of references to sources outside the field of psychology. This is the reciprocal of the ratio of Psychology References to Inter-Disciplinary Reference, and provides an index of thematic diversity.
c. Method to Introduction ratio: ratio of references cited in the method section of the document to the ones cited in the theoretical introduction.
d. Foreign language Percentage: the reciprocal of the ratio between the number of references in Spanish to the total number of references.
e. Index of Qualitative Variation: applied to the 5 document types (sources) which can be cited in the theses: books, journals, other theses, personal communications and web documents. This index is a measure of variability in the use of sources. See appendix for details.
f. Percentage of Locally Available References: is the proportion of documents cited which were obtained through the local Universidad Católica Andrés Bello library.

Time-based indicators

Lag, computed as the difference between the average date of the references and the year of completion of the thesis. Computed for average total reference list (total), books & Journal Articles.

Analysis

In order to minimize the consequences of a possible underestimation (or sampling variation) of the absolute quantities in the reports, *robust* statistics and confidence intervals are reported for all variables. Robust confidence intervals were obtained using the *bootstrap* re-sampling technique (see Appendix).

While the description of the sample is the main focus of the analysis, an additional multivariate analysis was conducted to highlight the main features of the dataset. Both factor and cluster analysis techniques were explored, and the cluster analysis model was used to identify the different bibliometric profiles in the cohorts studied and to create a classification of the thesis according to their multivariate differences in the bibliometric quantities.

Ethical considerations

None of the authors or advising professors' names were used in the creation of the databases, preser-

ving their privacy. Once the quality assurance operations were performed, each thesis was identified by an arbitrary numerical value, without the possibility of backward identification. The results of this analysis have no consequences whatsoever for the authors of the thesis or their advising professors, as all of the thesis analyzed had been completed and approved years before performing this analysis. Data analysis is conducted through a statistical, group-oriented procedure, without focusing on any particular document. All the data used is part of the records kept by the thesis advisory board, and no additional information was required from any individual.

Results

Table 1 shows the raw quantities for the total number of 125 documents. Reporting these absolute values was deemed important, since they offer a more accurate picture of the depth of the literature survey of each thesis than that provided by the relative indexes. One key feature in this table is the difference between mean and trimmed-mean (t-m) estimates. The trimmed mean excludes the top and bottom 5% of the distribution, focusing in the central 90%. The consistent difference between the mean and the trimmed mean indicates that a more robust statistic should be used to describe these distributions, and in consequence, the analysis is focused on median values. Bootstrap 99% confidence intervals provide robust upper and lower bound values for both the mean and the median. Another feature in Table 1 is that low frequency quantities (number of theses and web documents) have a large dispersion, with coefficient of variation (CV) values above 100.

A notable exception to the spread of these raw variables is the total number of references. This is especially evident in the Quartile data, which show that 50% of the theses have reference lists ranging between 41 and 61 items in length.

These raw bibliometric variables showed no discernible trends throughout the six cohorts studied (Table 2). Most of the differences between the years are conditioned by the different number of thesis within each cohort: 21, 19, 27, 17,15 and 26, for years 2001, 2002, 2003, 2004, 2006 and 2007, respectively. Differentiating the series for number of papers and number of books, the average absolute yearly change for the number of papers is 4.7 and 2.7 for the number of books. Building an error bar based upon inter-cohort variations yields a va-

Table 1. Descriptive statistics of the raw number of reference items by source, section, lenguage, field and local availability

Source	M[a]	CI(99%)[b]	t-m[c]	SD[d]	CV[e]	Md[f]	CI(99%)[g]	Q1[h]	Q3[i]	SIQ[j]	MIN	MAX
Journal Articles	31.12	27.9-34.66	29.67	16.57	53	27	25-30	21	37	8	5	108
Books	17.31	15.5-19.01	16.97	8.5	49	16	13-19	10	23	6.5	2	39
Thesis	4.64	3.84-5.46	4.32	3.94	85	4	3-5	2	6	2	0	19
Open Web	2.76	2.09-3.48	2.39	3.36	122	1	1-3	0	4	2	0	16
Personal Comm.	0.52	0.32-0.77	0.35	1.09	210	0	0-0	0	1	0.5	0	8
Total References	56.35	51.94-60.91	54.6	21.62	38	52	47-55	41	66	12.5	27	147
				Section								
Introduction	47.36	43.33-51.44	45.8	19.49	41	43	39-48	34	57	11.5	16	115
Methods	12.03	10.51-13.53	11.5	7.21	60	11	9-12	7	15	4	1	48
				Language, Field and Availability								
Spanish	23.02	20.46-25.52	22.23	12.33	54	21	18-24	14	29	7.5	0	68
Psychology	44.67	40.35-49.38	43.35	21.5	48	40	37-45	32	51.5	9.75	1	116
Locally Available	8.24	6.24-10.64	6.78	10.6	129	6	4-7	2	10	4	0	66

[a] Mean. [b] Bootstrap Confidence Intervals for the Mean. [c] 90% Trimmed Mean [d] Standard Deviation. [e]. Coefficient of Variation. [f] Median. [g] Bootstrap Confidence Intervals for the Median. [h] Fist Quartile. [i] Third Quartile. [j] Semi-interquartile range

riation range which falls within the bootstrap confidence intervals presented in Table 1, indicating that the point estimates and their robust confidence intervals can be used to take into account inter-cohort variations. Estimating time-based trends is difficult given the variability in the conditions for each academic year and the low number of time points, since theses are completed on a yearly basis. Using the 3 main quantities as an example shows that the number of papers cited and the total number of references are fluctuating quantities, while the number of books is in steady decline from 2001 to 2006, but it rises again in 2007. All these factors and results, combined with the lack of data for 2005 lead to the conclusion that there are not clear yearly trends in the raw quantities observed.

The bibliometric indexes (see Tables 3), as compared with the raw quantities, exhibit a more robust mean, since the difference with the trimmed mean is less noticeable than for the raw quantities. Journal articles proved to be by far the most widely used source for the sample as a whole, followed by

Table 2. Descriptive statistics of the raw number of reference items by year

Source		Median	Q1[a]	Q3[b]	SIQ[c]	MIN	MAX	N
Journal Articles								
	2001	27	18.0	33.0	7.50	15	73	21
	2002	31	26.0	54.5	14.25	9	108	19
	2003	23	18.0	29.5	5.75	5	50	27
	2004	29	22.0	41.0	9.50	6	55	17
	2006	27	23.0	29.0	3.00	12	40	15
	2007	31	22.5	41.5	9.50	18	72	26
Books								
	2001	21	13.0	27.0	7.00	5	38	21
	2002	19	11.5	28.5	8.50	2	39	19
	2003	17	10.5	23.0	6.25	5	34	27
	2004	15	12.0	26.0	7.00	6	30	17
	2006	11	9.0	14.0	2.50	7	31	15
	2007	15	9.8	19.0	4.63	4	32	26
Undergraduate Thesis								
	2001	5	2.0	8.0	3.00	0	14	21
	2002	4	1.5	6.0	2.25	0	19	19
	2003	2	1.0	4.0	1.50	0	12	27
	2004	4	3.0	6.0	1.50	0	15	17
	2006	5	4.0	6.0	1.00	2	13	15
	2007	4	2.0	7.0	2.50	0	15	26
Open Web								
	2001	1	0.0	3.0	1.50	0	14	21
	2002	1	0.0	1.0	0.50	0	12	19
	2003	2	0.0	3.0	1.50	0	10	27
	2004	5	2.0	7.0	2.50	0	12	17
	2006	1	1.0	5.5	2.25	0	10	15
	2007	1	0.0	3.8	1.88	0	16	26
Locally Available References								
	2001	8	3.0	12.0	4.50	0	61	21
	2002	5	0.5	11.5	5.50	0	44	19
	2003	6	1.5	9.0	3.75	0	29	27
	2004	7	5.0	15.0	5.00	1	66	17
	2006	3	0.0	7.0	3.50	0	17	15
	2007	5	2.3	8.0	2.88	0	24	26

Continues…

Source		Median	Q1[a]	Q3[b]	SIQ[c]	MIN	MAX	N
Total No. References								
	2001	55	47.0	69.0	11.00	30	110	21
	2002	64	40.5	94.0	26.75	28	147	19
	2003	45	36.0	55.0	9.50	27	78	27
	2004	54	44.0	74.0	15.00	36	109	17
	2006	44	40.5	51.5	5.50	33	64	15
	2007	53	44.5	66.0	10.75	38	119	26

[a] Coefficient of Variation. [b] First Quartile. [c] Third Quartile.

books, and with other thesis and open web references being only marginally used. The scarce use of open web references is widespread across the sample, with half the theses analyzed citing two web references or less. Nonetheless, as a whole, the sample exhibited moderate levels of source diversity, as reflected in the mean of the Qualitative Variation Index. Another widespread characteristic of the literature reviews on this sample of thesis is the heavy reliance on foreign language sources and material not available at the local library. Indeed, three fourths of the thesis included at least 48% of foreign language sources (most commonly, sources in English, data not included) and no more than 20% of references were locally available. Cross-disciplinary referencing, on the other hand, was highly variable throughout the sample, with minimum and maximum values encompassing the full range of the index (0 to 100%). A similar, but less extreme, pattern may be observed for the Methods to Intro ratio: while the average minimum and maximum values of this index differ widely, three fourths of the thesis used less than one reference in the Methods section for every three references cited in the introduction. This points to the presence of a few outlier thesis with an emphasis on the method section, rather than on the substantive content. As to the average age of the references used, the books cited were in general older than the Journal articles, while the average age of the references remained highly homogeneous across the sample, as shown by the high similarity of the mean, median and quartiles.

As with the raw bibliometric values, the indexes failed to exhibit discernible trends along the cohorts (see Tables 4 and 5). The median values for all indexes fall roughly within the confidence intervals for the sample as a whole.

Table 3. Descriptive statistics of the Bibliometric Indexes

Indexes	Mean	CI(99%)[a]	t-M[b]	SD[c]	CV[d]	Median	CI(99%)[e]	Q1[f]	Q3[g]	SIQ[h]	MIN	MAX
% Journal Articles	55	51-58	55	16	29	55	51-59	44	66	11	14	93
% Books	32	29-35	31	14	44	31	26-35	22	40	9	7	81
% Thesis	8	7-9	8	6	72	6	6-8	4	10	3	0	23
% Open Web	5	4-6	4	6	121	2	2-4	0	8	4	0	30
% Locally Available	14	11-17	12	15	108	10	7-12	4	20	8	0	88
% Foreign Language	59	55-62	59	18	31	61	56-66	48	72	12	9	100
%Cross-Discip. Ref.	21	14-26	20	28	139	14	8-22	3	37	17	0	100
% Qualitative Variation	68	65-70	68	13	20	70	67-71	61	76	8	17	97
Methods to Intro Ratio	0.28	0.24-0.35	0.26	0.28	99	0.23	0.21-0.29	0.17	0.33	0.08	0.02	3.00
Lag Books	12.20	11.48-12.93	12.12	3.53	29	12	11-13	10	14	2	5	22
Lag Journal Articles	9.96	9.09-10.93	9.65	4.47	45	9	9-10	7	13	3	3	31
Total Lag	10.17	9.51-10.89	10.01	3.32	33	10	9-11	8	12	2	4	22

[a] Bootstrap Confidence Intervals for the Mean. [b] 90% Trimmed Mean [c] Standard Deviation. [d] Coefficient of Variation. [e] Bootstrap Confidence Intervals for the Median. [f] Fist Quartile. [g] Third Quartile. [h] Semi-interquartile range

Table 4. Descriptive statistics of the Bibliometric Indexes by year

Indexes	Median	Q1[a]	Q3[b]	SIQ[c]	MIN	MAX	N
% Journal Articles							
2001	50	36	56	10	27	77	21
2002	63	48	75	13	28	93	19
2003	53	43	60	9	14	79	27
2004	51	44	59	8	17	77	17
2006	58	51	61	5	28	76	15
2007	59	49	69	10	35	83	26
% Books							
2001	39	25	43	9	10	69	21
2002	28	17	37	10	7	56	19
2003	35	26	48	11	15	81	27
2004	34	22	36	7	12	58	17
2006	25	19	30	5	13	49	15
2007	29	19	34	7	9	53	26
% Locally Available							
2001	12	5	29	12	0	88	21
2002	8	2	13	6	0	43	19
2003	10	3	22	9	0	40	27
2004	11	9	26	8	2	83	17
2006	6	0	14	7	0	35	15
2007	9	4	13	4	0	26	26
% Foreign Language							
2001	61	47	71	12	20	79	21
2002	68	54	74	10	9	100	19
2003	64	47	73	13	15	87	27
2004	60	49	70	11	14	78	17
2006	60	50	69	9	21	75	15
2007	57	48	74	13	19	87	26
%Cross-Discip. Ref.							
2001	21	8	41	17	0	100	21
2002	5	0	22	11	0	47	19
2003	22	9	40	15	0	86	27
2004	11	5	33	14	0	91	17
2006	3	0	25	13	0	59	15
2007	20	5	43	19	0	99	26
% Qualitative Variation							
2001	71	63	79	8	49	92	21
2002	66	49	71	11	17	85	19
2003	68	65	72	4	40	97	27
2004	76	65	81	8	47	88	17
2006	70	70	75	3	48	90	15
2007	67	59	75	8	37	90	26
Methods to Intro Ratio							
2001	0.21	0.12	0.27	0.07	0.02	0.47	21
2002	0.19	0.11	0.23	0.06	0.08	0.45	19
2003	0.23	0.18	0.30	0.06	0.11	0.62	27
2004	0.29	0.21	0.44	0.12	0.09	0.60	17
2006	0.36	0.24	0.44	0.10	0.10	0.65	15
2007	0.30	0.20	0.33	0.06	0.07	3.00	26

[a] Coefficient of Variation. [b] First Quartile. [c] Third Quartile.

Table 5. Descriptive statistics of the age of references by year

Age of Reference	Median	Q1	Q3	SIQ	MIN	MAX	N
Lag Books							
2001	11	9	13	2	5	21	21
2002	14	11	17	3	8	22	19
2003	13	11	14	2	7	22	27
2004	11	8	14	3	6	19	17
2006	12	10	13	2	5	17	15
2007	12	11	15	2	7	19	26
Lag Journal Articles							
2001	11	7	13	3	3	31	21
2002	8	5	11	3	4	23	19
2003	9	8	13	3	4	23	27
2004	10	8	15	4	3	18	17
2006	9	7	10	2	4	14	15
2007	9	7	12	3	4	15	26
Total Lag							
2001	10	8	12	2	4	22	21
2002	10	8	12	2	6	21	19
2003	9	8	13	2	5	17	27
2004	10	8	12	2	4	14	17
2006	9	8	11	2	6	13	15
2007	9	8	13	2	5	17	26

Cluster Analysis

Partitioning cluster analysis results are presented in this section Table 6. Another way of approaching this data is to explore the covariance structure (i.e. a factor model), however, there are issues with the correlation matrix, as many indexes have low specific variances, being redundant for the covariance structure, however those indexes are important for practical purposes. Q-oriented cluster analysis resulted in a more reliable computational solution, allowing the inclusion of the main indexes, and is therefore presented as the multivariate model for this dataset. For the sake completeness, a correlation matrix is included in the Appendix. The interpretation of the cluster analysis results is supported by the correlation structure.

The number of clusters was determined by the balance in the distribution of the observations and the separation among clusters. A number of different algorithms were employed, and the best solution was the K-means algorithm, which uses Euclidean distances between points and Means as cluster centroids (Leisch, 2006). Among solutions ranging from 2 to 5 clusters, the 2 cluster model yielded a more even distribution across the clusters, and the highest level of inter-cluster separation (average distance between clusters=2.4). Also, the 2 cluster solution was the most parsimonious.

The number of observations per cluster were 61 and 64, for clusters 1 and 2 respectively, representing 49 and 51% of the sample. This frequency count was, by far, the most evenly distributed of all the other cluster models.

Cluster 1 features higher proportion of papers (see Figure 1), minimum proportion of books, less variation in sources and content, larger proportion of foreign language references and shorter lags in reference date. Cluster 2 shows a higher proportion of books, a higher use of resources in the local library (Figure 3, and Table 6) and higher source variation.

The most noticeable differences are in the proportion of foreign language references, (22% more for cluster 1, see Figure 2), proportion of papers (21% more for cluster 1), and the proportion of books (17% more for cluster 2).

Table 6. Differences in Bibliometric Indexes for Clusters 1 and 2

	Mean		SD[a]		CV[b]		Median		Q1[c]		Q3[d]	
Index	C 1	C 2	C 1	C 2	C 1	C 2	C 1	C 2	C 1	C 2	C 1	C 2
% Journal Articles	67	42	9	10	13	24	66	45	60	37	74	50
% Books	22	41	8	12	36	29	23	40	15	34	28	49
% Thesis	7	9	5	6	75	67	6	7	3	4	10	13
% Open Web	3	7	4	7	131	103	2	4	0	0	5	10
% Locally Available	10	18	10	18	103	101	8	13	2	6	12	24
% Foreign Language	71	47	10	16	15	35	71	49	65	39	78	60
%Cross-Discip. Ref.	18	23	32	25	180	107	12	19	2	4	31	40
% Qualitative Variation*	59	76	11	9	19	12	62	76	52	71	68	81
Methods to Intro Ratio	0.29	0.27	0.37	0.15	127	55	0.26	0.23	0.16	0.18	0.31	0.35
Lag Books	11.6	12.8	3.5	3.5	30.0	28.0	12.0	13.0	9.0	11.0	13.0	14.3
Lag Journal Articles	9.7	10.3	4.1	4.8	42.0	47.0	9.0	10.0	7.0	7.0	12.0	13.0
Total Lag	9.9	10.4	3.3	3.3	33.0	32.0	9.0	10.0	8.0	8.0	12.0	13.0

[a] Standard Deviation. [b] Coefficient of Variation. [c] First Quartile. [d] Third Quartile.

* See Appendix for details

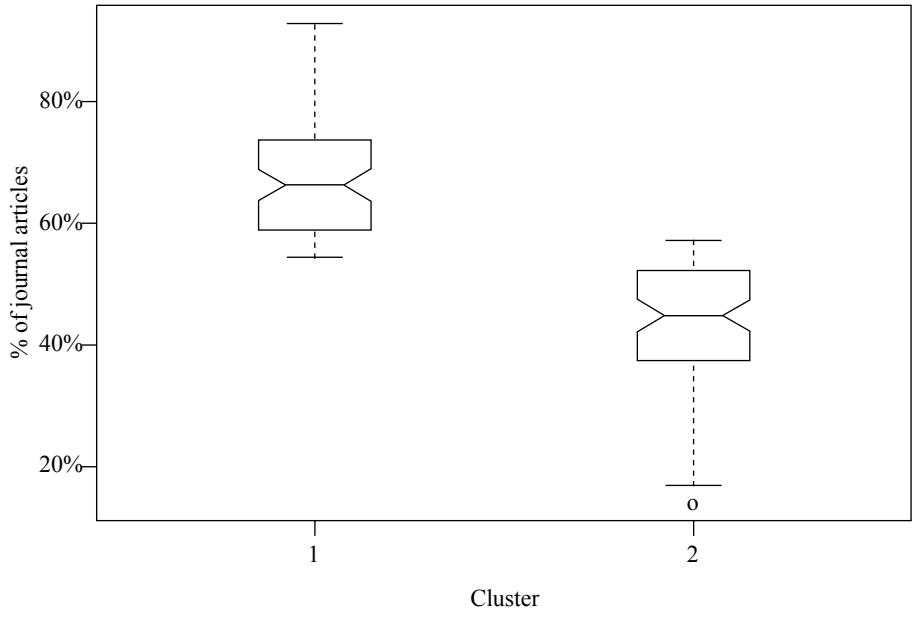

Figure 1. Boxplot of the Ratio of Papers by Cluster. Bottom, mid and top of the boxes represent Q1, Q2 (median) and Q3, respectively. Notches round the middle point are proportional to the standard error of the median. Lines beyond the boxes extend toward the 5 and 95 percentile of the distribution. Points beyond these lines are considered outliers.

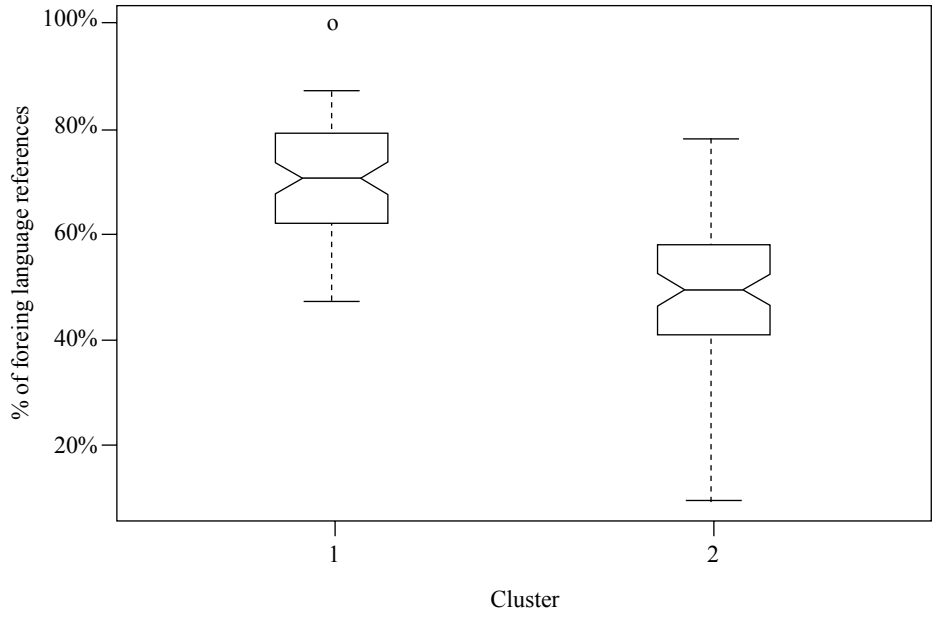

Figure 1. Boxplot of the proportion of foreing language references by cluster

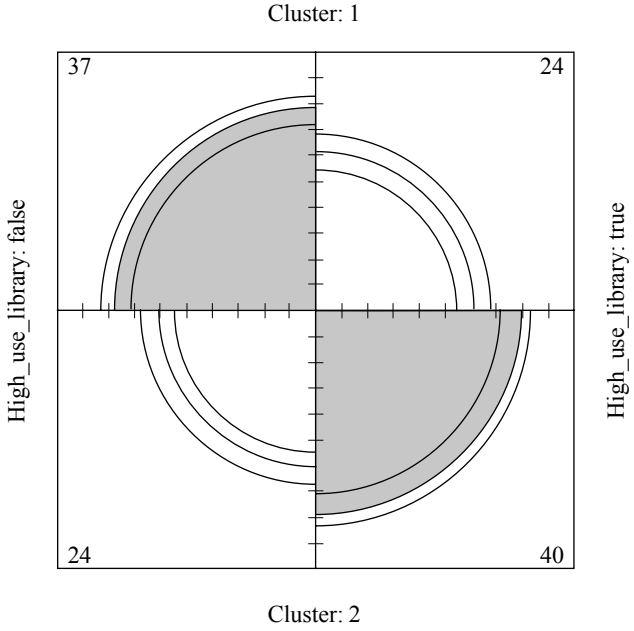

Figure 3. Fourfold plot of Cluster by high use of local library. Segment separation represent cell value departure from change expectation (Null hypothesis). Central ring represents the log-odds ratio contribution of each cell and outer rings represent 95% confidence interval for the log-odds ratio. Counts for each cell are indicated near each segment.

Discussion

The shows the typical bibliographic profile of the licentiate thesis studied. The bootstrap confidence intervals allow considering the estimates stable enough to serve as reference values for policymaking about the standards for undergraduate thesis.

The extension of the literature review made in each thesis (as portrayed by the median number of references used) is difficult to gauge, due to the scarcity of bibliometric research on undergraduate scientific production. Bibliometric analysis of term papers by American undergraduate students report a median bibliography length of 10 reference items for the social sciences (Mill, 2008), while median reference list length for Master´s thesis at Iowa State University and Virginia Tech are reported to be of 33.5 and 30.5 items, respectively (Kushkowski, 2005). The reference length for Doctoral Dissertations reported in the literature exhibits a wide range: the median length for Iowa State University was reported to be of 69.5 items (Kushkowski, 2005), that of Virginia Tech of 76 items (Kushkowski, 2005), 91 for the Institut Químic de Sarriá (IQS) of Barcelona, Spain(Vallmitjana & Sabate, 2008), and 105 for Doctoral Dissertations in Education at the University of Minnesota (Haycock, 2004). Comparable bibliometric data from Latin American undergraduate thesis are even more scarce: Aguilar, López-López, Barreto, Rey, Rodríguez and Vargas (2007) report an average reference list of 32 items for undergraduate thesis in Organizational Psychology in a Colombian University. Clearly then, the depth of the literature reviews of the present dataset do not match those of doctoral dissertations, but are considerably longer than that of undergraduate term papers, undergraduate psychology thesis in other Latin American universities and Master´s thesis in several American universities.

As to the predominant references sources (journal articles, 67% and books, 22%), the present results match more closely the patterns found in undergraduate papers in the sciences than in the humanities or the social sciences. In his bibliometric analysis of papers from 64 courses in an American Mid-Atlantic college, Mill (2008) reported that for courses in the Humanities, references to books compromised 60.7% of citations, with journal articles made up just 24.5% of all references. For the Sciences, on the other hand, these proportions were practically reversed, with journal articles making up 66.2% of citations, while books accounted for only 17.3% of all references. Social Sciences showed a similar, but less extreme pattern, with citations to articles reaching 46.7 % and references to journals 25.2%. Similar differences between the use of journals for doctoral dissertations in the Biological and Social Sciences have been reported by Kushkowski, Parsons & Weise (2003).

The scarce use open web citations (median=5%) and heavy reliance on foreign language literature (median=59%) may be explained by the emphasis of most research methods courses and the thesis supervisory system at UCAB on peer-reviewed publications and use of foreign language references, mainly in English. Nevertheless, similarly low use of open web sources has been reported for doctoral dissertations in American universities (Kushkowski, 2005), undergraduate papers (Mill, 2008), and papers written by psychology faculty (Schaffer, 2004). Rather than constituting a new source of content, the World Wide Web seems to be providing more efficient access to traditional references, at least as far as the scholarly literature is concerned.

Regarding the age of the citations, in general, the average lag was of 10 years in the present sample, with 9 years for journal articles and 12 years for books. This can be considered a high standard for undergraduate theses, as those numbers are typical of journal-quality publications in social/behavioral sciences. The kind of source used, as well as the relatively low age of the references, can be considered an effect of the thesis supervisory system employed during most of the time period studied (Dillon & Malott, 1981; Gant, Dillon, & Malott, 1980; Robles, Csoban-Mirka & Vargas-Irwin, 2009).

One final outstanding feature of the present data is the very low availability of references in the local library. While first-world bibliometric studies show local availability figures ranging between 62 and 89% (Mill, 2008; Schaffer, 2004), for the present sample, on average, only 14% of references were

obtained locally. Although no similar data from other third world countries were available for comparison, the more meager resources may result in less duplication of collections between academic institutions, thus forcing students to use libraries from several universities in order to carry out thorough literature reviews.

Finally, in spite of the general orientation of the theses towards a natural science scholarly style, the multivariate cluster analysis of the present data shows that the distribution of the thesis along the continuum proposed by Hargens (2000) is not homogeneous, but rather allows for the characterization of two distinct groups: one which resembles more closely the natural science literature, featuring a higher proportion of references to journal articles, a smaller proportion of references to books, less variation in sources and content, larger proportion of foreign language references and shorter lags in reference dates, and a second group of thesis, closer to the scholarly style of the humanities, with higher proportion of citations to books, a higher use of resources in the local library and higher source variation. This heterogeneity in the scholarly style of undergraduate thesis in psychology may vary well respond to the diversity of the content areas within the field. The stability of the raw quantities, as well as the indexes, alongside the alignment of the results towards two kinds of bibliographic profiles (that of the natural sciences or the humanities) may provide valuable insights for policymaking and standards of bibliographic profiles for licentiate theses.

References

Adams, J. (1998). Benchmarking international research (Editorial Material). *Nature, 396*(6712), 615-618.

Aguilar, M.C., López López, W., Barreto, I., Rey, Z.B., Rodríguez, C. & Vargas, E.C. (2007). Análisis bibliométrico de los trabajos de grado del área organizacional de la Facultad de Psicología de la Universidad Santo Tomás. *Diversitas, 3*(2), 317-334.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E. & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics, 3*(4), 273-289.

Ardila, R. (1986). *La psicología en América Latina: pasado, presente y futuro*. México DF: Siglo XXI Editores.

Buchanan, A.L. & Herubel, J. (1994). Profiling PHD dissertation bibliographies - serials and collection development in political-science. *Behavioral & Social Sciences Librarian, 13*(1), 1-10.

Dillon, M.J. & Malott, R.W. (1981). Supervising masters theses and doctoral dissertations. *Teaching of Psychology, 8*(4), 195-202.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68*, 589-599.

Gant, G.D., Dillon, M.J. & Malott, R.W. (1980). A behavioral system for supervising undergraduate research. *Teaching of Psychology, 7*(2), 89-92.

Haberman, S.J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association, 77*, 568-580.

Hargens, L.L. (2000). Using the literature: reference networks, reference contexts and the social structure of scholarship. *American Sociological Review, 65*(6), 846-865.

Haycock, L.A. (2004). Citation analysis of education dissertations for collection development. *Library Resources & Technical Services, 48*(2), 102-106.

Hirsch, J.E. (2007). Does the h-index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America, 104*(49), 19193-19198.

Klassen, R.M. & Kuzucu, E. (2009). Academic procrastination and motivation of adolescents in Turkey. *Educational Psychology, 29*(1), 69-81.

Klassen, R.M., Krawchuk, L.L. & Rajani, S. (2008). Academic procrastination of undergraduates: low self-efficacy to self-regulate predicts higher levels of procrastination. *Contemporary Educational Psychology, 33*(4), 915-931.

Kushkowski, J.D. (2005). Web citation by graduate students: a comparison of print and electronic theses. *Portal-Libraries and the Academy, 5*(2), 259-276.

Kushkowski, J.D., Parsons, K.A. & Wiese, W.H. (2003). Master's and doctoral thesis citations: analysis and trends of a longitudinal study. *Portal-Libraries and the Academy, 3*(3), 459-479.

Leiding, R. (2005). Using citation checking of undergraduate honors thesis bibliographies to evaluate library collections. *College & Research Libraries, 66*(5), 417-429.

Leisch, F.A. (2006). Toolbox for K-Centroids Cluster Analysis. Computational statistics and data analysis, *51*(2), 526-544.

Mathur, V.P. & Sharma, A. (2009). Impact factor and other standardized measures of journal citation: a perspective. *Indian J Dent Res, 20*(1), 81-85.

Mill, D.H. (2008). Undergraduate information resource choices. *College & Research Libraries, 69*(4), 342-355.

Mooney, C.Z. & Duval, R.D. (1993). *Bootstrapping: a nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: a review. *Scientrometrics, 66*(1), 81-100.

Nederhof, A.J., Zwaan, R.A., Debruin, R.E. & Dekker, P.J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioral-sciences. A comparative-study. *Scientometrics, 15*(5-6), 423-435.

Reynolds, H.T. (1984). *Analysis of nominal data* (2nd. ed.). Newbury Park, CA: Sage.

Robles, J.R. (1996). PRS polytomous item generation-simulation according to the common-factor model. *Applied Psychological Measurement, 20*, 140.

Robles, J.R., Csoban-Mirka, E. & Vargas-Irwin, C. (2009). Análisis cuantitativo de la dinámica individual de trabajos de grado de Psicología. *Suma Psicológica*, *16*, 51-68.

Rosario, P., Costa, M., Núñez, J.C., González-Pienda, J., Solano, P. & Valle, A. (2009). Academic procrastination: associations with personal, school and family variables. *Spanish Journal of Psychology, 12*(1), 118-127.

Rousseeuw, P.J. & Van Driessen, K. (1999). A Fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*, 212-223.

Schaffer, T. (2004). Psychology citations revisited: behavioral research in the age of electronic resources. *Journal of Academic Librarianship, 30*(5), 354-360.

Seo, E.H. (2008). Self-efficacy as a mediator in the relationship between self-oriented perfectionism and academic procrastination. *Social Behavior and Personality, 36*(6), 753-763.

Thompson, D.F., Callen, E.C., & Nahata, M.C. (2009). New indices in scholarship assessment. *American Journal of Pharmaceutical Education, 73*(6), art. 111.

Vallmitjana, N. & Sabate, L.G. (2008). Citation analysis of Ph.D. dissertation references as a tool for collection management in an academic chemistry library. *College & Research Libraries, 69*, 72-81.

Van der Hulst, M. & Jansen, E. (2002). Effects of curriculum organization on study progress in engineering studies. *Higher Education, 43*(4), 489-506.

Walters, W.H. (2008). Journal prices, book acquisitions and sustainable college library collections. *College & Research Libraries, 69*(6), 576-586.

Appendix

Qualitative dispersion

One of the main approaches for the quantification of the dispersion of nominal variables is based on a *concentration* measure. Concentration is defined as:

$C = 1 - \sum p_i^2$

where p_i is the probability of the nominal variable *Y* to assume the value of the i-th discrete category (see Haberman , 1982, for a review). The Index of Qualitative Variation is defined as

IQV= [k/k-1] * *C*

where k is the total numbers of nominal categories (Reynolds, 1984). IQV is supposed to take any possible value from 0 to 1. Applied to the bibliographic source data, k is the number of source types (books, journals, web, etc), and the index measures the amount of variability of the citations across different sources. Lower values indicate concentration in few sources and higher values reflect dispersion in the use of a variety of sources.

Bootstrap confidence intervals for means and medians

Due to the peculiar nature of the group of documents studied, with the added factor of its multi-cohort nature, it may be a stretch to make strong distributional assumptions about the data, and in consequence, estimation of standard errors and confidence intervals built by standard parametric methods may be inappropriate. Moreover, given that the raw quantities and indexes are of special interest to aid in policymaking of thesis supervision and regulation, robust confidence intervals were considered an important analysis requirement.

One way to build robust, non-parametric confidence intervals is a re-sampling method called bootstrap. This method produces robust confidence intervals based on a maximum likelihood approximation of the Empirical Distribution Function (*EDF*) (Efron, 1981). The bootstrap procedure employed to build confidence intervals for bibliometric indexes can be summarized in the following steps:

1. Definition of the *EDF*. In this case, each observation has a probability p=1/n.
2. Sample with replacement, Q samples of size n, according to the *EDF.*
3. For each sample, compute and store statistics for simulated sample, creating the vector $S^{<j>}$ of length Q, with statistical results for each of the *j* statistics (In this case, *j*=2, mean and median).
4. Use contents of vector $S^{<j>}$ as the sampling distribution of the statistical value. Estimate percentile values corresponding to the upper and lower bounds of the confidence interval for each of the *j* statistics.

There are other ways to build bootstrap confidence intervals, such as the normal method and the bias-corrected method. However, the percentile method was considered straightforward and less loaded with distributional assumptions (Mooney & Duval, 1993).

Bootstrapping is a computer-intensive procedure, especially when involve location parameters like the median, and the chosen percentile method to estimate the confidence intervals require a large number of simulated samples (Q=10000) in this case, and intensive use of random number generators. To implement this procedure in a feasible way, a parallel algorithm was designed and programmed by the authors, to take advantage of a multi-core CPU, and an object-oriented implementation of random number generators was used (J. R. Robles, 1996), adapted for parallel computing in this case.

Correlation matrix for indexes and lags

The correlation matrix R shows standard Pearson moment-product correlations above the diagonal, and robust estimates below the diagonal, obtained via the Minimum Covariance Determinant (MCD) procedure. MCD is used to obtain robust covariance (and associated correlation) matrix estimate, by recombination of the observed data until the determinant of the matrix |R| is minimized. The obtained covariance matrix considered a robust estimate, based upon a selected subset of the observed data. The algorithm used in this case is an incremental random re-sampling procedure called FastMCD (Rousseeuw, 1999).

The main diagonal shows the lower-bound estimate of communality for each variable (Square Multiple Correlation, SMC). The SMC for each variable is computed using linear algebra algorithm based upon Cholesky decomposition of the correlation matrix. Results are equivalent to obtaining the SMC regressing each variable on the rest of the variables in the matrix.

Recepción: 11 de noviembre de 2009
Aceptación: 5 de marzo de 2010