International Journal of Internet Science
2008, 3 (1), 7–33

ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

Objectivity, Reliability, and Validity of Search Engine Count Estimates

Dietmar Janetzko

National College of Ireland, Dublin, Ireland

Abstract: Count estimates ("hits") provided by Web search engines have received much attention as a yardstick to measure a variety of phenomena of interest as diverse as, e.g., language statistics, popularity of authors, or similarity between words. Common to these activities is the intention to use Web search engines not only for search but for ad hoc measurement. Using search engine count estimates (SECEs) in this way means that a phenomenon of interest, e.g., the popularity of an author, is conceived of as a measurand, and SECEs are taken to be its quantitative measures. However, the data quality of SECEs has not yet been studied systematically, and concerns have been raised against the use of this kind of data. This article examines the data quality of SECEs focusing on classical goodness criteria, i.e., objectivity, reliability, and validity. The results of a series of studies indicate that with the exception of Boolean queries that use disjunction or negation objectivity as well as test-retest reliability and parallel-test reliability of SECEs is good for most types of browsers and search engines examined. Estimation of validity required model development (all-subsets regression) revealing satisfying results by using an explorative approach to feature selection. The findings are discussed in the light of previous objections and perspectives for using Web search count estimates are delineated.

Keywords: Data quality, goodness criteria, Web mining, search engines, search engine counts

Introduction

Search engines do not only search. Google, Yahoo, MSN, and other major search engines act as global gateways of information interchange unprecedented in their scope and depth. In recent years, sampling from search engines and using the number of results (hits, or search engine count estimates, SECEs) returned has become an active area of research in information science, database design, linguistics, and social sciences (Lawrence & Giles, 1998). Seen from a more general point of view, this research follows one of two rationales differing in emphasis of their preferred research methods. First, work centering on the *development of algorithms* intended to generate estimates of search engine measures (e.g., on search engine coverage). Secondly, work pivoting around *search engine statistics* that uses the number of results returned by search engines to analyze phenomena not directly related to the Internet (e.g., word similarity). Work following the first strand was initially motivated by the goal to estimate the index size of search engines (Bharat & Broder, 1998) and has since diversified considerably both with respect to the main approaches used like random sampling (Bharat & Broder, 1998; Bar-Yossef & Gurevich, 2006; Schuster & Schill, 2007) or random walk (Henzinger, Heydon, Mitzenmacher, & Najork, 1999; Rusmevichientong, Pennock, Lawrence, & Giles, 2001) and the scope of estimations the proposed algorithms strive to achieve (e.g., corpus size, overlap of different search engines, index freshness, density of duplicates, query hits). Work associated with the second strand is focused on statistical analysis of the number of

Address correspondence to Dietmar Janetzko, National College of Ireland, Mayor Street, IFSC, Dublin 1, Phone +353 1 4498 610, Fax: +353 1 4060 559, dietmarja@gmail.com

SECEs returned upon entering a query to a search engine. This data is used as a yardstick to measure a variety of phenomena of interest as diverse as language statistics (Pullum, 2004; Krug, 2006; Hundt, Biewer, & Nesselhauf, 2007), popularity of authors (Bagrow & ben-Avraham, 2005), similarity between words (Cilibrasi & Vitanyi, 2007) or mappings between concept hierarchies (Gligorov, Kate, Aleksovski, & Harmelen, 2007). Common to these activities is the intention to use Web search engines for ad hoc measurement. This work indicates that using query hits is beginning to gain acceptance as a kind of data that facilitates scientific studies though a number of case studies have raised concerns about using query hits as data (Bar-Ilan, 2001; Rousseau, 1999; Wouters, Hellsten, & Leydesdorff, 2004).

However, notwithstanding the far-reaching conclusions of work taking a sceptical stance towards SECEs, its methodological status is debatable. Bar-Ilan (1999) conducted a case study that used only the search phrase *informetrics OR informetric*. The focus of this analysis was not on the research results (SECEs) but on the results (URLs). Data was collected in one month intervals during a five months period in 1998 and an additional search round in 1999 using the six largest search engines at that time (Altavista, Excite, Hotbot, Infoseek, Lycos, and Northern Light). Her work indicates that search engines do not only discover new URLs, they also forget URLs they knew before even though they continue to exist. This resampling of the search engine index may or may not lead to changes to the overall number of SECEs. It is, however, unclear whether and to what degree the results ought to be attributed to mechanisms on the side of the search engines or to the disjunction used in the query. Bagrow and ben-Avraham (2005) examinded larger samples of concepts – called populations by the authors – but were mainly concerned with distribution fitting. In a more recent study, Wouters et al. (2004) made use of the search string frankenfood* OR (frankenstein AND food*) to collect SECEs from Altavista. The authors also collected SECEs from Google. But since Google has different query format requirements the original search string had to be split into three different ones. The pooled outcome was then compared with the results obtained from Altavista. Over a period of two years search was carried out about ten times in irregular intervals. In sum, previous work on the data quality of SECEs does not consider the conceptual framework used in the empirical sciences to assess data.

A systematic study that assesses the data quality in terms of the goodness criteria of data quality, i.e., objectivity, reliability, and validity is still missing. Work that adheres to the first strand above does no address the question of data quality of SECEs either since its focus is on the development on algorithms that approximate parameters of search engines.

The work presented in this paper follows the second of the two strands of research mentioned above, i.e., the methodological examination of search engine result data (SECEs). It addresses the research question whether and to what degree SECEs comply with standard goodness criteria of data quality, i.e., *objectivity, reliability*, and *validity* (e.g., Carmines & Zeller, 1991; Odom & Morrow, 2006). All studies presented in this paper make use of external evaluations to examine SECEs. This means that the overall process of generating SECEs (e.g., generation, size, and organization of search engine indices, algorithms employed to estimate SECEs by the search engines studies, etc.) is treated as a black box and only the goodness of SECEs is examined. This paper is organized as follows. The first part of this paper gives an outline of Web search count estimates and the second part presents a sequence of studies conducted to examine objectivity, reliability, and validity of SECEs. The article concludes by highlighting the potential and limits of SECEs and gives an outline of possible future application of this type of data.

Search Engine Count Estimates

When entering a query term many search engines do not only provide a list of Web links but also a figure that indicates the number of documents found that satisfy the query launched. Contrary to a widespread belief, this figure does not express the exact count of pages that relate to the search query (“hits”) but an estimate of this number that may or may not be subject to fluctuations. For instance, the number of SECEs may differ between different pages of the numbered result list many search engines present at the bottom of the page that responds to a query. Likewise, repeatedly clicking on the search button of a search engine may or may not produce different counts for SECEs. Given the volatility of this type of data the term *search engine count estimates* (SECEs) is proposed in this paper and preferred over alternative terms like *results*, *hits* or *page counts*. The latter terms suggest a high level of correctness that seems to be inappropriate with respect to Web search count estimates.

Eliciting Search Engine Count Estimates

Search engine count estimates can be collected in a number of ways. The most straightforward one is to enter one or several search words into a search engine (e.g., Google, Yahoo, MSN, Altavista) that returns SECEs. But using the Web interface of a search engine to elicit SECEs is laborious and time-consuming. It can hardly be

employed if SECEs of a large number of search terms are to be studied. The same is true if SECEs are to be elicited at regular intervals, e.g., on a daily basis. There are alternatives to the Web interface available that allow the researcher to automate searching on the Web and eliciting SECEs. Google and Yahoo offer some support to automate the extraction of SECEs (Mayr & Tosques, 2005; Google, Inc, 2008; Yahoo, Inc, 2008), but Google stopped issuing new keys required to make use of this API at the end of 2006. A number of different programing language (Java, C#) or scripting languages (Perl, Python) can be used to query Google via a computer program or script. For instance, Landers (2008) developed a Python script that connects to the Google application programming interface (API), which can be taken to trigger Google search operations. An alternative to the Google API is to make use of a script that starts a console-based browser, e.g., Lynx, w3m, or Links, which launches a search engine query, reads the results and saves the SECEs. Several Perl scripts have been presented in news groups that support this approach (Anonymous, 2005).

Objectivity, Reliability, and Validity

The quality or goodness of data can be described on different levels. Traditionally, the goodness of data is conceived of in terms of objectivity, reliability, and validity. In disciplines which often have to deal with noisy data, e.g., psychology, social sciences, economics, and medicine the concepts of reliability, objectivity, and validity play a vital role and have motivated a rich methodological literature (e.g., Carmines & Zeller, 1991; Odom & Morrow, 2006; Gruijter & Kamp, 2007). The study of SECEs can certainly profit from this work. Still, a methodology tailored to the analysis of SECEs is required. The goal of this paper is to contribute to its development. Seen from a statistical vantage point, objectivity, reliability, and validity are based on correlations. Coefficients for each of these goodness criteria range between 0 and 1 indicating that achievement of them is a matter of degree. But clearly, correlation coefficients underlying objectivity, reliability, and validity are only useful shortcuts, which should always be embedded into a more elaborate analysis that conveys a more complete picture of the patterns of the data studied.

Objectivity. Objectivity addresses the question if the data collection is independent of the persons involved in data collection and independent of the devices used. Consider the example of a thermometer used for measuring the temperature. If thermometers of various vendors differed considerably data collection that makes use of them would not qualify as objective. What may affect objectivity when collecting SECEs? We can rule out the possibility that objectivity of collecting SECEs depends on the person carrying out the Web search. However, the multitude of different technical set-up or different locations of the person or the program conducting the search (e.g., search done in Europe vs. search done in China) may or may not influence search results. To examine objectivity, Study 3 examines the objectivity of browsers as the major client-side software used to collect SECEs.

Reliability. How consistent are SECEs when assessed at different points in time or when elicited by using alternative approaches to data collection? This question addresses the reliability of SECEs. Reliability informs about the extent to which the repeated use of a measure leads to consistent, i.e., the same or comparable values. Only if data is highly reliable may we safely conclude that changes of magnitude do not reflect fluctuations or errors but changes of the phenomenon under study. For this reason, a high reliability of data is indispensable in science and engineering. Reliability can be estimated in several alternative ways, each of which casts a light on a particular aspect of reliability. Among the most common types of reliability are test-retest reliability and parallel-test reliability (Gruijter & Kamp, 2007). Test-retest reliability informs about the consistency of longitudinal data and parallel-test reliability gives an account on cross-sectional data. More specifically, while test-retest reliability is a measure of the consistency of results from one point in time to another, parallel-test reliability expresses the consistency of results obtained via different data collection methods. Study 4 examines test-retest reliability of SECEs and Study 5 is an investigation of parallel-test reliability of SECEs.

Validity. Validity concerns the degree to which a measure expresses a phenomenon it is taken to reflect. Validity studies examine whether and to what degree empirical evidence, variables that already have a well-understood meaning or both are in agreement with a finding the validity of which ought to be determined. To conceive of SECEs in such a way that validity becomes an issue reflects a more recent tendency in search engine usage. Validity is at stake whenever SECEs are considered to mean something other than themselves. This is true when SECEs provided by Google are used as a numeric indicator of popularity of persons, software, etc. (Bagrow & ben-Avraham, 2005). In areas where calculating validity has become part of a standard scientific procedure, e.g., IQ testing, routines have been developed to study validity. As yet, however, no established procedure exists to examine the validity of SECEs. In line with more recent considerations of validity (Gruijter & Kamp, 2007) the ultimate goal is to work toward a theory of the phenomena to be studied instead of breaking down validity into many different types the relationship among which remains often unclear.

Search engine count estimates are summary scores influenced by many variables. To examine their validity SECEs have to be decomposed and traced back to the variables that contribute to their magnitude. In itself, a SECE is a number that does not reveal the variables that lead to its magnitude. Consider the example of using the number of SECEs of a city, e.g., Boston. What does this figure mean? Is it possible to validate it successfully by relating it to the size of the population in the city? Or do SECEs of cities indicate a high intensity of Internet-related activities, a high crime rate, a high concentration of companies that advertise a lot, the frequent usage of a particular city's name in marketing activities, especially high or low prices of apartments or a large number of cultural activities in a city? These are only some factors that may contribute to the number of SECEs of cities. In short, when estimating the validity of SECEs of a target concept, e.g., city, variables need to be identified that correlate highly with this data. In statistics and data mining, this is a task known as the feature or predictor selection problem (Guyon & Elisseeff, 2003). In Study 6 validity of SECEs was examined by using statistical techniques (regression analysis) to select predictors.

Studies

Overview of Studies. The studies presented in this paper examined whether and to what degree SECEs meet the goodness criteria of objectivity, reliability, and validity strived for in scientific investigations. The first two studies were preparatory in nature. Their results were required to examine the goodness of SECEs in the remaining studies. The goal of Study 1 was to select concepts required to carry out subsequent studies. Study 2 addressed the question of whether Boolean queries lead to sound magnitudes of SECEs. Study 3 pivots around the objectivity of SECEs. SECEs were collected via different browsers and the consistency between the results obtained was determined. Reliability of SECEs was the central topic of studies 4 and 5. In Study 4, reliability of SECEs was analyzed on the basis of SECEs obtained at different points in time. Study 5 examined reliability by comparing the SECEs of different search engines. Finally, Study 6 provides an example of the way validity of SECEs can be examined. Examination of validity is exemplified by studying the popularity of American cities on the basis of other general concepts assumed to be involved with the popularity of a city. The latter includes media coverage effects and effects relating the popularity of a city to topics like science, economy, crime, or culture and their associated SECEs. The validity study explores both the theoretical relationships between the more general concepts involved and the empirical relationships between the concepts and their observable indicators or measures (SECEs). The results obtained are interpreted with reference to the clarification of the construct validity of the measure (SECEs).

Query Terms. In studies 1 – 5 SECEs were examined on the basis of one-tuple (one word) index terms taken from the online version of Encyclopædia Britannica (Britannica, 2008). The online version of Encyclopædia Britannica covers 284,128 index terms (4th of April 2008). The majority of the index terms are n-tuples. An example of a n-tuple to be found in Encyclopædia Britannica is the following phrase

Zweites Deutsches Fernsehen (German television station): see ZDF.

By contrast, the number of one-tuple index terms, e.g., *anachronism*, is much smaller. For the following reasons only the comparatively small number of one-tuple index terms were chosen to elicit SECEs. Firstly, they were intended to be used in simple search queries. This means, while the Boolean operator AND was taken into consideration for some queries (studies 2 and 6) all other studies made use of simple queries without Boolean operators. Secondly, index terms should facilitate an unequivocal search. A number of n-tuples among the index terms are ambiguous when used in a Web search query. For instance, many n-tuples include a translation, abbreviation or cross-references to other index terms which makes them ambiguous when used in a Web search query. The one-tuple index terms chosen were mostly nouns and adjectives. They cover a broad spectrum of thematic fields like science, geography, history, religion, art, culture, and everyday life. It can be assumed that the index terms range from high to low frequency words. One-tuples of Encyclopædia Britannica include concepts that seem to be widespread and in common use, e.g., *city*, but also concepts that appear to be used only rarely, e.g., *ahgareseh* or *analvos*. After selecting all 825 one-tuple index terms found in the online version of Encyclopædia Britannica, 6 concepts had to be discarded since they almost certainly prevent unequivocal search. Among those was the concept *adobe* (clay) which was expected to lead to confusion. Often having several meanings, abbreviations such as *AWACS, FiOS,TIROS and TiVo* were not considered either. Discarding the index terms mentioned lead to a set of 819 one-tuples that were used in studies 1, 4 and 5 (see Appendix A). Query terms in Study 6 were 25 names of American cities and concepts hypothesized to have an impact on the number of SECEs of cities (e.g., *crime, culture, politics, science, CNN, New York Times, Washington Post, Internet*).

Study 1 – Baseline Data

The goal of this study was to establish baseline data on the magnitude of SECEs for the 819 concepts that were selected from the online version of Encyclopædia Britannica and used as query terms (see Appendix A). Searches were conducted using Google because a previous study indicates that in contrast to other search engines Google updates its index on a daily basis (Lewandowski, Wahlig, & Meyer-Bautor, 2005).

Method

A Google search of 819 one-tuple index terms from Encyclopædia Britannica was conducted. The text-based browser Lynx along with a Perl script was employed to carry out all search operations and to process the results. The operating system used was Intel Mac OS X 10.5.

Results

The Zipf distribution has been repeatedly found in studies of word frequencies (Zipf, 1932) including randomly generated texts (Li, 1992). Zipf 's law reflects the fact that a relatively small number of concepts is used very often, while an abundance of other concepts are rarely used. This phenomenon has also been found in the sample of query terms used to study SECEs. In this study, a ranking among the query terms was established on the basis of the observed magnitude of the SECEs. With some exceptions the observed frequencies of the 819 query terms studied followed Zipf 's distribution (Figure 1). In general, the SECE of a more frequent query term was considerably higher than the SECE of query term on the next lower rank position which gave the resulting distribution a long tail to the right (power law).

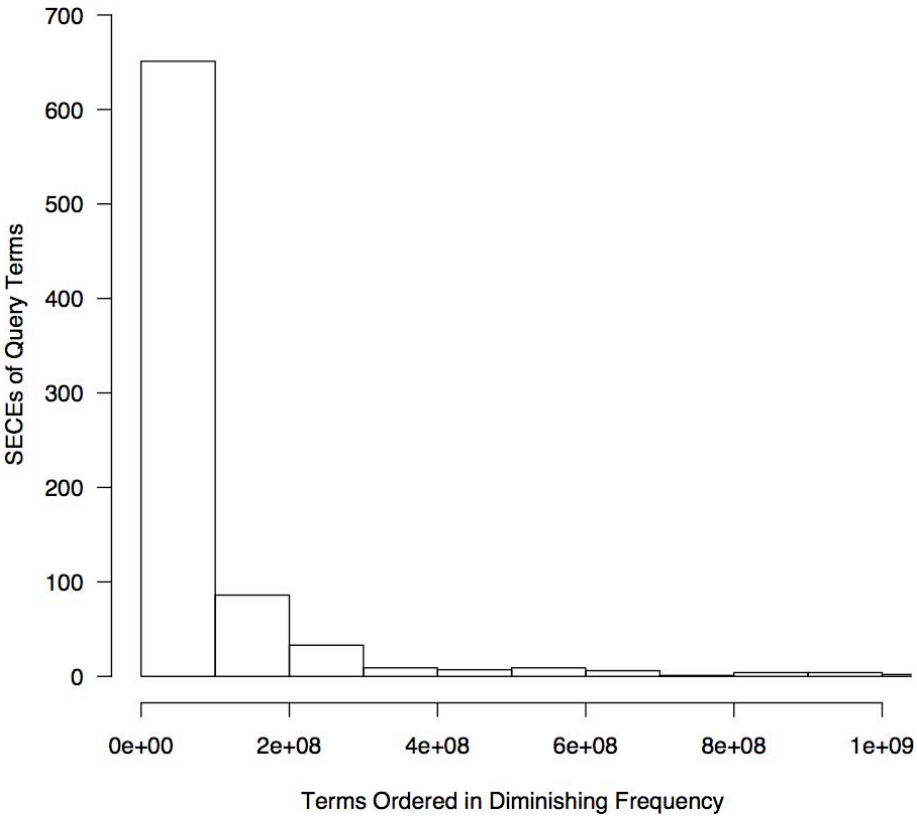


Figure 1. Frequency spectrum of the query terms studied (n = 819 query terms).

On the basis of the rank order of all 819 concepts gleaned from the online version of Encyclopædia Britannica a subset of concepts was drawn that comprised 45 query terms (concepts) differing in magnitude of SECEs (high, medium, low). The rank position of the concepts within the overall group of concepts was used as the criterion to allocate concepts to the groups of high, medium and low scoring concepts (cf. Table 1). This set of concepts will be used in Study 2 to examine the effect of Boolean queries on SECEs and in Study 3 to study the objectivity of SECEs. The majority of the low frequency concepts generated 1-4 digit SECEs, medium frequency concepts had typically about 7-digit SECEs, and most of the high frequency concepts lead to 9-10 SECEs. Given the power law distribution of the 819 concepts shown in Figure 1 the low and medium frequency concepts occupy a position at the left end of the distribution while the high frequency concepts are to be found at the far right end of the distribution.

The concepts presented in Table 1 suggest a negative correlation between word length and SECEs. In fact, it is a well-established finding that the length of words in a large language corpus and its usage are negatively correlated (Zipf, 1932; Whaley, 1978). In line with these findings, in the overall set of 819 concepts word length and SECEs were also found to be negatively correlated (Pearson's product-moment correlation, $r = -.2$, $p < .001$). Taken together, both the frequency spectrum of SECEs (Figure 1) and the negative correlation between word length and SECEs indicate that SECEs follow the rules of language usage described by Zipf and others.

Table 1
Study 1 – Concepts Generating High, Medium, and Low Magnitudes of SECEs (n = 819 Query Terms)

High Magnitude		Medium Magnitude		Low Magnitude	
Concept	Rank	Concept	Rank	Concept	Rank
art	001	headwear	406	gemmulation	806
computer	002	airfield	407	toponomastics	807
health	003	geophysics	408	waxplant	808
travel	004	bottling	409	trihexaflexagon	809
history	005	antidepressant	410	gibberfish	810
France	006	thermodynamics	411	cobiron	811
technology	007	apprenticeship	412	clinograde	812
food	008	heresy	413	ballistospore	813
hotel	009	hairdresser	414	Ceravix	814
entertainment	010	teapot	415	analvos	815
Canada	011	gastronomy	416	cataclastite	816
water	012	earphone	417	alguacile	817
education	013	handcuffs	418	avacchedakata	818
war	014	fortnight	419	anirmoksa	819
China	015	typewriter	420	ahgareseh	819

Discussion

The concepts gleaned from Encyclopædia Britannica were shown to cover a broad spectrum differing strongly in frequency and word familiarity. As expected, concepts like *art, computer* and *health* were found to generate a high number of SECEs while the concepts *avacchedakata, anirmoksa* and *ahgareseh* – arcane-looking to the uninitiated – were observed to lead to a low number of SECEs.

Study 2 – Boolean Search Queries

A number of search engines facilitate Boolean search queries, which allow users to search in a more focused way. Previous discussions on the data quality of SECEs indicate that search via Google using Boolean operators lead to a number of SECEs that contradict Boolean logic (Liberman, 2005). Study 2 investigates whether Boolean search queries generate results in line with Boolean logic. Let a, b, and c be one-tuple concepts used as query terms. Let a signify a search engine query that makes use of concept a. The expression $|a|$ denotes the number of SECEs returned when launching a search a. Boolean queries and the number of SECEs returned are formalized accordingly. For instance, $|(a \wedge b)|$ is taken to mean the number of SECEs upon launching a conjunctive search engine query of a and b. If Boolean operators work correctly then a number of minimum requirements on the number of SECEs should be fulfilled. The focus in this study is on three basic Boolean query types: disjunction, conjunction and negation (see Table 2). The requirements spelled out in this table specify intervals for SECEs when launching Boolean queries. The interval expectation derived from these requirements can then be compared with SECEs obtained when launching corresponding Boolean queries. Thus, the requirements facilitate to diagnose whether SECEs comply with Boolean logic. The Boolean statements underlying this study can be described as follows.

Disjunction. When launching a disjunctive search query that makes use of two concepts a and b, e.g., Tango OR Internet, the resulting number of SECEs should be at least as high as the number of SECEs of the higher scoring concept (SECEs for Internet), but equal or lower than the sum of SECEs of a and b searched in isolation (SECEs for Tango + SECEs for Internet).

Conjunction. When launching a conjunctive search query that makes use of two concepts a and b, e.g., Tango AND Internet, the resulting number of SECEs should be the same as or lower than the number of SECEs of the lower scoring concept (SECEs for Tango) but equal or higher than zero.

Negation. Using a negation in a search query that involves two concepts *a* and *b*, e.g., Internet -Tango, the resulting number of SECEs should be the same or smaller than the number of SECEs of the higher scoring concept (SECEs for Internet) and smaller than the difference of SECEs of the concepts *b* and *a* searched in isolation (SECEs for Internet -SECEs for Tango).

Table 2
Number of SECEs expected in Boolean Queries

Disjunction	$\max(\|a\|, \|b\|)$	$\leq$	$\|(a \vee b)\|$	$\leq$	$\|a\| + \|b\|$
Conjunction	$\min(\|a\|, \|b\|)$	$\geq$	$\|(a \wedge b)\|$	$\geq$	0
Negation	$\|b\|$	$\geq$	$\|(\neg a, b)\|$	$\leq$	$\|b\| - \|a\|$

Method

It cannot be ruled out that the estimation procedures used by search engines to generate SECEs produce different results depending on the frequency of the query concepts. In order to detect possible frequency-dependent effects, the terms taken to launch Boolean queries were 45 one-tuple concepts that were shown to generate high, medium, and low numbers of SECEs in Study 1 (Table 1). Each of the 3 × 15 concepts was deployed in Google queries that made use of the Boolean operators AND, OR, negation,[1] and the operand *Internet*. This query term was chosen because it guarantees that a high number of SECEs is returned. In addition, the 45 one-tuples were used without Boolean operators in order to obtain baseline data. Thus, the number of queries in Study 2 was (3 + 1) × 15 and ranged from Internet, art, Internet AND art, Internet OR art, Internet -art to ahgareseh, Internet, Internet AND ahgareseh, Internet OR ahgareseh and Internet -ahgareseh. Conjunctive and disjunctive queries used the format *operand operator operand* (infix notation). The Boolean operator (AND OR) was written in capital letters. Negation was conducted by using the format *operand -operand*. Note that any deviation from this format (e.g., using small letters for Boolean operators in disjunctive or conjunctive queries, changing the sequence of operands or operators) will lead to different numbers for SECEs. Search was carried out manually by using Firefox 3.04 running on Intel Mac OS X 10.5 as the operating system. Cookies were allowed. No Google account was used and no attempt was made to camouflage the IP.

Results

Disjunction. Boolean queries that made use of a disjunction resulted in magnitudes of SECEs that were not in line with Boolean logic. Complying with principles of Boolean logic the disjunction of two concepts *a* and *b* never lead to a smaller frequency of either *a* or *b*. Anomalies became apparent, however, among low frequency concepts. Consider the example of the query term *cobiron*. This word had a SECE of 2,130 (Google, 15th of December 2008). The word *Internet* generated a SECE of 2,020,000,000. The SECE for the disjunction Internet OR cobiron was 1,910,000,000. This result is not possible if rules of Boolean logic would have been applied.

Conjunction. Boolean queries that made use of a conjunction resulted in SECEs the frequencies of which comply with principles of Boolean logic. In line with Boolean logic the SECEs of all concepts were always below the number of SECEs obtained for the lower scoring concept.

Negation. When launching a boolean query that made use of a negation the number of SECEs returned contradicts principles of Boolean logic. Complying with principles of Boolean logic a query that makes use of a negation, e.g., *–a b* never lead to a magnitude of SECEs higher than the SECEs of *b*. However, concepts of all groups when used in a Negation lead to biased results. Consider the example of the query term *ahgareseh*. This word had SECEs of 39 (Google, 15th of December 2008). The word *Internet* had a SECE of 2,020,000,000. The SECE for the negation (Internet -ahgareseh) was 1,910,000,000. This outcome would not have occurred if principles of Boolean logic had been applied.

Discussion

Previous concerns about the data quality of Google SECEs elicited via Boolean queries (Liberman, 2005) could partially be confirmed. Study 2 found distorted results for all Boolean queries except for conjunctions of type *a* AND *b*. The finding that Boolean Google queries of type *a* AND *b* provide sound results in line with Boolean logic is important because this type of query is required for search related to the study of validity of SECEs in Study 6.

[1] In Google queries, negation is expressed via '-' directly attached to a search a search term, e.g., '-Tango'. If 'NOT' is used Google will return documents that contain the word 'not'.

Study 3 – Objectivity

Does the magnitude of SECEs depend on the Web browser used? This question is crucial because it relates to the objectivity of this measure. Study 3 examines the objectivity of SECEs by examining whether different Web browsers yield comparable figures of SECEs.

Method

An experimental procedure was chosen to investigate potential effects of browser types on the number of SECEs returned. With the exception of Internet Explorer version 6.0 and version 7.05, which ran under Windows XP, all browsers used Intel Mac OS X 10.5 as the operating system. The search engine was the same across all queries of this study (Google), while the browser types were varied. The experimental procedure was indispensable to trace back potential differences to the kind of browser used for searches. The experimental procedure rested on the assumption that the search engine, i.e., its index size, algorithms, etc. did not change drastically while the study was conducted. Therefore, all data recordings of Study 3 were carried out on the same day (4-24-2008). As in studies 2, 3 and 4, query terms used were 45 one-tuple concepts that were shown to generate high, medium and low numbers of SECEs in Study 1. To examine potential browser dependency five different browsers were selected for search engine queries. These were Firefox 2.0.04, Internet Explorer version 6.0 and version 7.05, Safari 3.1.1 and Lynx 2.8.6. The Internet Explorer version 6.0 and version 7.05 and Firefox 2.0.04 are currently (June 2008) used by more than 90% of all Internet users (W3schools, 2008). In addition, Lynx, a text-based browser and Safari, the major browser for Apple computers, were used to examine potential browser dependency of SECEs. Search was carried out manually with the browsers mentioned above.

Results

Spearman's rank correlation was calculated to estimate the closeness of association of SECEs among all pairs of search engines. The results are shown in Table 3.

Table 3
Study 3 – Objectivity of Collecting SECEs (Consistency between SECEs Povided by Different Web Browsers Engines expressed by Spearman's Rank Correlation, n = 45 Query Terms)

	Lynx 2.8.6	Firefox 2.0.0.4	Safari 3.1.1	IE 6.0
Firefox 2.0.0.4	0.91***			
Safari 3.1.1	0.97***	0.89***		
IE 6.0	0.91***	1.00***	0.89***	
IE 7.05	0.92***	1.00***	0.88***	1.00***

***$p < .001$

Discussion

Overall, Study 3 revealed that correlations between SECEs of all browsers and thus the objectivity of SECEs were found to be high. It could be shown that collecting SECEs via Firefox or Internet Explorer lead to high SECEs and the results obtained from these browsers were almost identical. Other browsers considered in this study reached correlations on lower levels. The finding that the magnitude of SECEs collected by Lynx does not completely match up with SECEs collected by other browser types is noteworthy because it means that automated collection of SECEs via Lynx is not fully in line with SECEs gleaned by way of other browsers.

Study 4 – Test-Retest Reliability

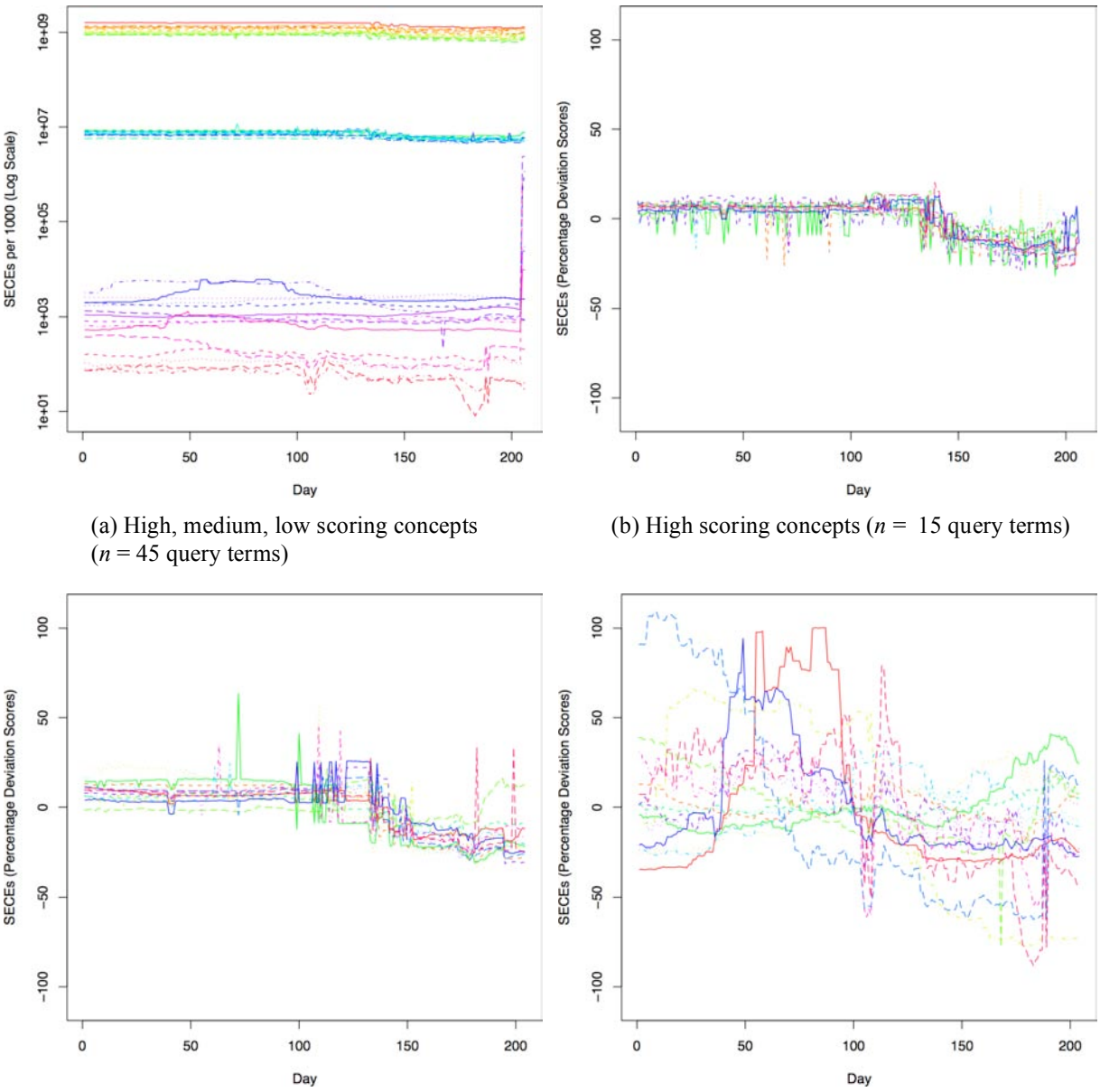
How consistent is the number of SECEs of one search engine when assessed at different points in time? The question addresses the test-retest reliability of SECEs. We may reasonably expect that the test-retest reliability of concepts differs strongly with the kind of concepts used in a search engine query. For instance, in a situation where the world economy is about to slide into a recession the word *recession* can be expected to generate more SECEs than at a time when the economy is doing well. Other concepts, e.g., the article *the* may not be affected by global events that provoke a sizable media resonance but by changes of the corpus used by a search engine. Likewise, the test-retest reliability of SECEs might differ with the overall frequency of concepts because the estimation procedure used by search engines may be affected more (or less) by a high number of entries in the search engine corpus. This study investigates test-retest reliability of SECEs using a breakdown of concepts associated with a high, medium, and low number of SECEs as identified in Study 1. The motivation for this breakdown is two-fold. Firstly, previous studies suggest that the accuracy of the estimation algorithms used by search engines to generate SECEs depends on the frequency of associated documents. For instance, Bagrow and ben-Avraham (2005) assumed that SECEs might be more accurate for rare concepts that generate a small

number of SECEs. Secondly, the breakdown used is likely to uncover results that are useful for subsequent research (e.g, in the sense that SECEs of rare concepts are more reliable or vice versa).

Method

As in studies 2 and 3, 45 concepts differing in frequency (high, medium low, Table 1) were used as query terms to elicit SECEs. Data recording started in May 22, 2008. None of the concepts could be expected to vary considerably as a consequence of events on a global scale.[2] Google SECEs were collected on a daily basis over a period of 205 days. The text-based browser Lynx along with a Perl script was employed to carry out all search operations and to process the results. The operating system used was Intel Mac OS X 10.5. For all concepts the variation of the magnitude of SECEs in terms of the percentage of deviation from the mean were calculated.

Results



(a) High, medium, low scoring concepts (n = 45 query terms)

(b) High scoring concepts (n = 15 query terms)

(c) Medium scoring concepts (n = 15 query terms)

(d) Low scoring concepts (n = 15 query terms)

Figure 2. Study 4 – development of Google search engine estimates over a period of 205 days (n = 3 × 15 query terms).

[2] An exception is the concept *China* because the Olympic Games 2008 took place in China from August 8-24. However, China was big in the media long before the start of the Olympic games (e.g., because of the row between China and Tibet that became more expressed in the months preceding the start of the Olympic games. This is why no clear-cut sudden peak is visible in the data.

Figure 2 illustrates data collection of SECEs over a period of 205 days. To give an overview of SECEs of all three groups studied, i.e., concepts generating a high, a medium and a low magnitude of SECEs, the first subfigure maps the results (SECEs) on a log scale. Each of three remaining subfigures 2b – 2d centers SECEs at an average of zero and express results as percentage deviation scores each In this way, the results become comparable between the three groups of concepts considered. Subfigure 2a indicates that none of the concepts shows drastic changes in terms of its magnitude of SECEs. This means, while there are variations in terms of the number of SECEs each of the concepts stays within the boundaries of concepts that score on a high, medium, or low frequency level of SECEs. In other words, the intra-concept variation of SECEs was relatively low. By contrast, inter-concept differences of SECEs were high. Each of the three groups of concepts showed different patterns of SECEs. In the group of high scoring concepts some concepts exhibited a periodic pattern of change of their SECE frequency and remained relatively stable across the 205 days the data recording was carried out. Starting with day 100 of the data recording, the concepts that yield a medium number of SECEs showed a higher level of variations pattern of SECEs. The largest variations of SECEs were found in the group of low scoring concepts, i.e. concepts that generate 1-3 digit SECEs. Figure 2 shows consistently lower magnitudes of SECEs around the day 140. This attenuation is clearly expressed in the groups of high and medium scoring concepts. In the group of medium scoring concepts the attenuation is preceded by an increase of noise in the data. The most plausible reason for the attenuation are internal re-organisation of Google's search index or associated procedures.

Tables B1 – B3 (see Appendix B) present an overview of the test-retest correlations. Starting with day one of data recording (May 22, 2008), correlations were calculated for every tenth day of data recording. In general, correlation scores were high, indicating a high degree of consistency among SECEs of concepts (query terms) in time. In line with the descriptive account of SECEs given in Figure 2 correlation scores were highest for concepts that trigger a high and – to a lesser degree – medium number of SECEs. For about three months correlation scores for high and medium scoring concepts remained high. After this time, correlation scores for medium scoring concepts deteriorated considerably. Concepts (query terms) that generated a low number of SECEs reached a high level of consistency only in the first 4 weeks of data recording.

Discussion

Search engines have to update their index in short intervals to keep up to date. For this reason, it is almost inevitable that SECEs vary in time. Still, the question is when variations occur and which magnitude they have. The major finding of this study is that among concepts that yield a high number of SECEs variability is lowest and reliability is highest. Long-term variability is higher among concepts with a medium number of SECEs. Among concepts that generate a low number of SECEs variability is highest and test-retest reliability is lowest. This finding contradicts the untested assumption of Bagrow and ben-Avraham (2005): *It seems reasonable to assume that very small counts are more accurate than larger ones* (Bagrow & ben-Avraham, 2005, p. 81). In fact, it seems more plausible that the estimation procedure misses out on concepts with a small number of entries in the search engine corpus, which results in a low test-retest reliability. The findings obtained in this study were shown on a descriptive level and they are reflected in the magnitude of test-retest reliability of concepts of all three groups examined.

Study 5 – Parallel-Test Reliability

How do SECEs compare to SECEs of the same concepts elicited via other search engines? This study investigates parallel-test reliability by using different search engines to collect SECEs and examines the degree of consistency obtained.

Method

819 concepts (one-tuple terms) identified on the basis of the index terms of Encyclopædia Britannica in Study 1 were employed as query terms. The search engines used to study parallel-test reliability of SECEs were Google, MSN, Live, Yahoo, Altavista, Excite, Infoseek, AOL, Savvysearch, Webcrawler, Alltheweb and Ask. SECEs were collected automatically by using the text-based browser Lynx and a Perl script to process the result pages returned by Lynx. The operating system used was Intel Mac OS X 10.5. Elicitation of SECEs and the analysis of parallel-test reliability of SECEs proceeded in four steps. Firstly, for each search engine the sum of all SECEs was calculated on the basis of the 819 concepts used as query terms. Secondly, bivariate Pearson product-moment correlation scores of SECEs of all search engines considered were calculated. Thirdly, each bivariate correlation score is the outcome of an independent study. This is why a meta-analysis (Hunter & Schmidt, 2004; Hartung, Knapp, & Sinha, 2008) was calculated that aggregated the results obtained. Fourthly, a regression model was set up, which facilitated prediction of SECEs. Though prediction via a regression analysis is not the centerpiece of reliability assessment it sheds light on patterns in the data studied. The regression analysis used

SECEs of Google as the criterion variable to be predicted by SECEs of other search engines considered in this study. The regression model taken to predict Google SECEs started with a full model that included SECEs of all other search engines considered in this study. This analysis was an exhaustive search in the space of all regression models conducted via all-subsets regression (Miller, 2002). This was followed by a least-square multiple regression analysis to actually predict SECEs and to examine model performance on the basis of selected models. Finally, regression diagnostics tests for collinearity, and heteroscedasticity were carried out.

Results

Sums of SECEs. Even a casual look at SECEs of different search engines reveals large differences. This was corroborated and expanded by calculating the sum of SECEs for all 819 concepts (one-tuple terms) for each of the search engines considered (see Figure 3). Excite and Webcrawler generated the lowest and Yahoo, Altavista, Infoseek, and Alltheweb the highest sum of SECEs across all 819 query terms used.

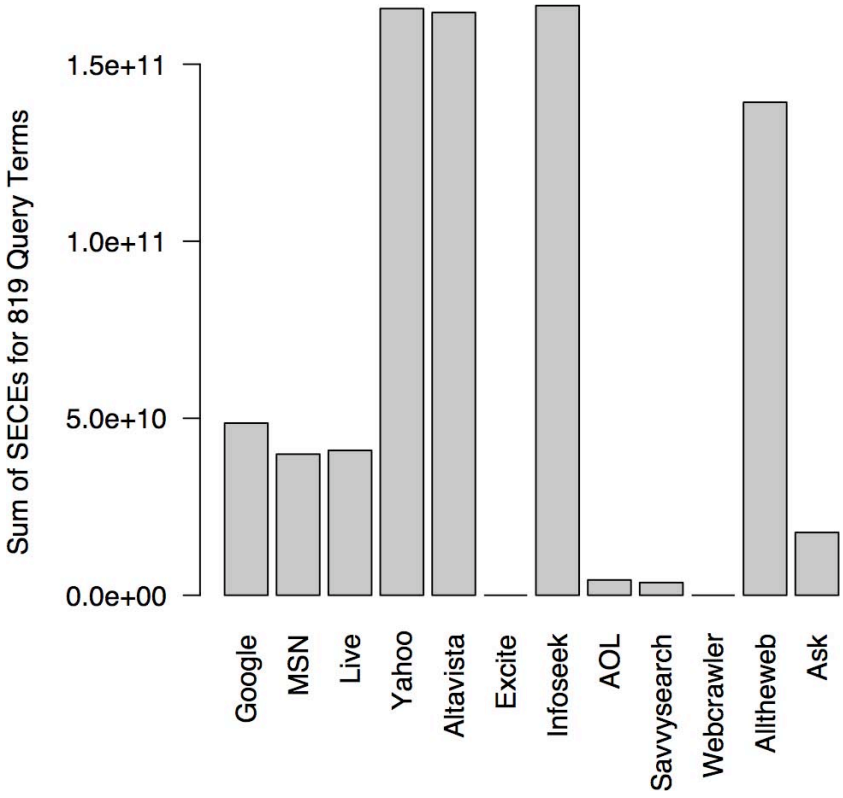


Figure 3. Study 5 – sums of search engine count estimates of different search engines (n = 819 query terms).

Bivariate Pearson's product-moment correlation of SECEs. The majority of the search engines studied revealed a high degree of parallel-test reliability as evidenced by high correlation scores (Table 4). SECEs of Google, MSN, Live, Yahoo, Altavista, Infoseek, AOL, Savvysearch, Alltheweb and Ask formed a cluster of high correlation scores with values exceeding 0.88. Among those, high correlation between Yahoo, Alltheweb, Altavista comes as no surprise since the latter two use Yahoo's search index. Close significant correlations were also observed between search engines that differed notably in their overall frequency figures of SECEs (e.g., Google and Altavista). Among the search engines studied only Excite and Webcrawler, i.e, the search engines with the lowest overall frequencies, did not join the cluster of close correlations (Figure 6). Different slopes in the subfigures of Figure 6 indicate that the average number of SECEs often differs between search engines. For instance, the flat slope in subfigures 4 to 6 of the second row of Figure 6 reflects the fact that MSN generates a comparatively low number of SECEs.

Meta-Analysis of Correlation Scores. The goal of meta-analyzing the parallel-test correlation scores was to compute mean correlation across studies corrected for sampling error (bare-bones meta-analysis, Hunter & Schmidt, 2004). In which way is the statistical tool of a meta-analysis used to analyze SECEs? Consider the situation when the parallel-test reliability of SECEs is calculated on the basis of SECEs obtained from two search engines. The result of this study is a product moment correlation coefficient that expresses the test-retest reliability. Clearly, there are more search engines available that facilitate similar studies, which also lead to results on parallel-test reliability. Willy-nilly, this leads to the question of how we deal with a possibly high number of findings that may or may not differ from each other. Meta-analysis has been designed to address this

question. It aggregates findings from different studies and facilitates a more general view on the phenomena studied. Moreover, meta-analysis is a powerful tool that supports an understanding of the true relationships or effects by accounting for artifacts and biases.

Ideally, all search engines considered for a study on the parallel-test reliability SECEs intended to be meta-analyzed should make use of the same set of query terms. Moreover, the search engines should be comparable with regard to the plausibility of their results. For instance, it would not make sense to consider a search engine that is known to reveal implausible results.[3] A meta-analysis will then aggregate the findings (coefficients of correlations) obtained for each study.

Table 4
Study on Parallel-Test Reliability of SECEs – Product Moment Correlation Scores between Search Engine Count Estimates of Different Search Engines, n = 819 Query Terms

	Google	MSN	Live	Yahoo	Altavista	Excite	Infoseek	AOL	Savvy-search	Web-crawler	Alltheweb
MSN	.81***										
Live	.83***	.99***									
Yahoo	.96***	.79***	.82***								
Altavista	.96***	.79***	.82***	1.00***							
Excite	-.020	-.010	-.01	-.02	-.02						
Infoseek	.96***	.79***	.82***	1.00***	1.00***	-.010					
AOL	.95***	.75***	.78***	.93***	.93***	-.02	.93***				
Savvysearch	.95***	.75***	.78***	.94***	.94***	-.02	.94***	1.00***			
Webcrawler	.21***	.17***	.18***	.19***	.19***	-.02	.19***	.25***	.24***		
Alltheweb	.93***	.76***	.80***	.99***	.99***	-.010	.99***	.90***	.91***	.19***	
Ask	.91***	.74***	.78***	.97***	.97***	-.010	.97***	.87***	.88***	.19***	.98***

***$p < .001$

The described procedure was administered to the coefficients of correlations found in this study (see Table 4) excluding the results for Excite and Webcrawler. The mean average correlation obtained on the basis of all highly correlating search engines was .877.

Predictions of SECEs. The first step of the prediction analysis was model selection, i.e., finding a suitable regression model that facilitated good prediction of SECEs. Model selection was carried out via all-subsets regression analysis (Miller, 2002). SECEs of Google were used as the criterion variable and SECEs of all other search engines considered (MSN, Live, Yahoo, Altavista, Infoseek, AOL, Savvysearch, Alltheweb, Ask) were the predictor variables. Excite and Webcrawler were again excluded. The model selection statistic used was adjusted R^2 (Lahiri, 2001). In addition to the predictor variables a noise variable ε derived from a random permutation of all SECEs observed was included in the regression analysis. This is a useful trick in model selection via regression analysis that helps select models and predictor variables (Flack & Chang, 1987). When inspecting the results every model that included the variable ε as a predictor – thereby attributing explanatory strength to a meaningless variable – was not considered.

The outcome of all-subsets regression is shown in Figure 4. This figure shows eight models arranged in eight horizontal rows ranging from low (bottom) to high (top) model performance. In Figure 4, model performance is indicated via the ordinate scale and also by different shades of gray used to present each single model. Light gray signals indicate comparatively low model performance and dark gray or black expresses comparatively high model performance. The most parsimonious model (shown in gray at the bottom of Figure 4) rests on one predictor variable (Altavista, adjusted $R^2 = .92$). This means that this model accounts for 92% of the variance in the criterion variable. Note that one of the models with the highest model fit uses 9 predictors ($R^2 = .96$). This applies to the model shown in the topmost row of Figure 4. However, since this model includes ε as a predictor variable it can safely be considered as meaningless. What is strived for in statistical model selection is an optimal balance between parsimony as evidenced by the number of predictors or parameters and a good model performance expressed via the goodness of fit statistic chosen (Ockham's razor, Zellner, Keuzenkamp & McAleer, 2001). This is why the most parsimonious model that included only one predictor (SECEs of Altavista) while at the same time providing a good model fit was selected (adjusted $R^2 = .92$). It makes use of one predictor to achieve a model performance that is only slightly worse than that of the 6 best performing models indicated by the 6 upper rows of Figure 4.

[3] For instance, a search engine like Webcrawler that returns 2 (!) hits for the search key *Obama* on the day preceding the presidential election of the USA (November 3, 2008) should not be considered.

For illustrative purposes, a multi-predictor model that included SECEs of Live, Yahoo, Altavista, Infoseek, AOL, Savvysearch, Alltheweb as predictors was also examined in the next step. The second step of prediction analysis was conducted to actually predict the criterion variable. A multiple regression analysis was calculated on the basis of a parsimonious one-predictor model with SECEs of Altavista as the only predictor (adjusted $R^2 = .92$) and a not-so-parsimonious multi-predictor model that used SECEs of Live, Yahoo, Altavista, Infoseek, AOL, Savvysearch, Alltheweb as predictor variables (adjusted $R^2 = .96$).

The performance of both models in terms of predicting SECEs of Google is illustrated in Figure 5. The comparison of both subfigures in this Figure shows that even the slim one-predictor model predicts Google SECEs very well. The performance of this model is only slightly improved if the number of predictors is increased (subfigure 5b). For this reason, the parsimonious one-predictor model is clearly the model of choice to predict Google SECEs.

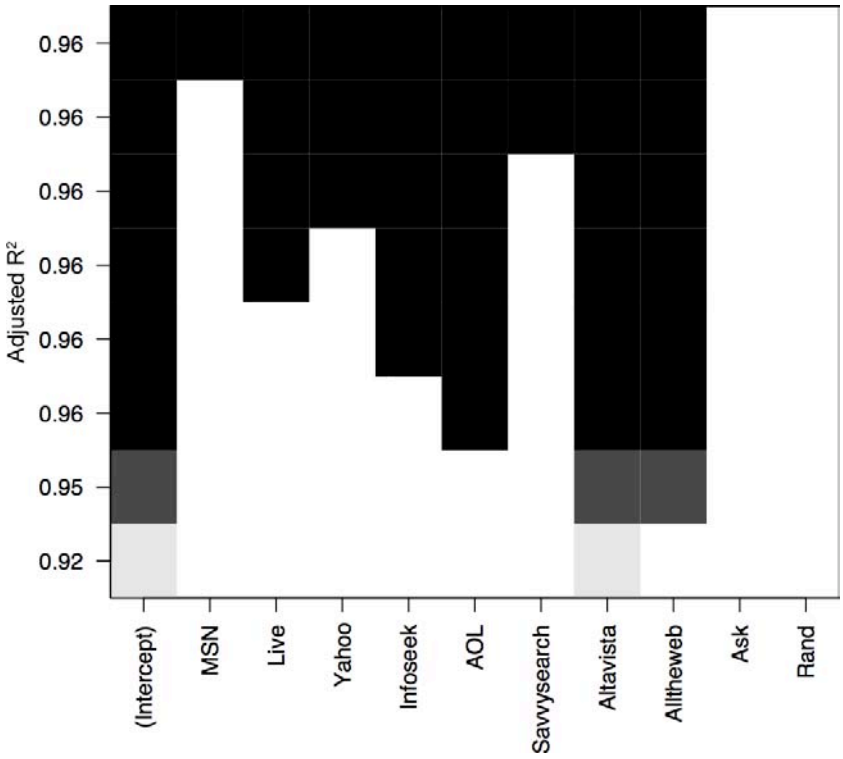


Figure 4. Study 5 – predicting Google SECEs via SECEs of other search engines (model selection, n = 819 query terms).

Regression Diagnostics. As expected, tests of the assumptions of regression analysis (regression diagnostics) showed that in the multi-predictor model collinearity was high among the predictors indicating strong correlations among each other. The Breusch-Pagan test revealed heteroscedasticity both for the one-predictor and the multi-predictor model. A more detailed analysis showed that in both models error variance was in fact smallest in the lower range of the criterion variable. This means, among low values of the criterion variable there was a tendency towards smaller and more comparable prediction errors. Correcting for heteroscedasticity by using heteroscedasticity-corrected covariance matrices (HCCM) to test the statistical significance of predictors (White, 1980; Long & Ervin, 2000) revealed that in the one-predictor model, Altavista SECEs were indeed significant ($t = 3.63$, $p <. 0001$). In the multi-predictor model, again Altavista SECEs ($t = 16.72$, $p < .0001$) and Alltheweb SECEs ($t = –2.43$, $p < .05$) reached the preset level of significance.

To test for collinearity of the multi-predictor model, variance inflation factors (VIF) were calculated (Fox, 1997). VIF factor scores for all predictors in the multi-predictor model were low for search engines Live and Yahoo (both VIF < 2) but high for all other predictors, which indicates strongly that most predictor variables in the multi-predictor model were confounded.

Discussion

Study 5 corroborated that among leading search engines the parallel-test reliability of SECEs is high. Quite expectedly, correlation between SECEs of the two Microsoft search engines MSN and Live was high. The finding that correlations between most other search engines was also very high comes as a surprise. Often, differences between competing search engines were not much higher than between cooperating search engines. A case in point is the high correlation between the search estimates of Google and those of Altavista, Yahoo and Alltheweb. Given that search engines differ in index size, index updating techniques, and the average number of SECEs returned it could not be expected that all search engines correlate highly. In fact, correlations between SECEs of both Excite and Webcrawler and all other search engines considered in this study were extremely low.

However, consistency among the SECEs of highly correlating search engines reached such a high level can use Altavista's SECEs to predict SECEs of most other search engines with a good degree of precision. This is an interesting finding since it indicates that research on SECEs need not suffer from an overreliance on a particular search engine.

The findings of Study 5 on the high parallel-test reliability of SECEs seem to contradict previous research that reported only a low overlap of results returned by search engines (Spink, Jansen, Blakely, & Koshman, 2006; dogpile.com, 2007). However, most of the studies that allegedly examined search engine result overlap confined themselves on the overlap of results found on the first page returned. On the assumption that many users take only the first result page of a search engine into consideration this confinement may be reasonable. But this does not justify conceiving the outcome as "result overlap". In contrast, Study 5 analyzed all results returned and did not limit data collection to the first page returned.

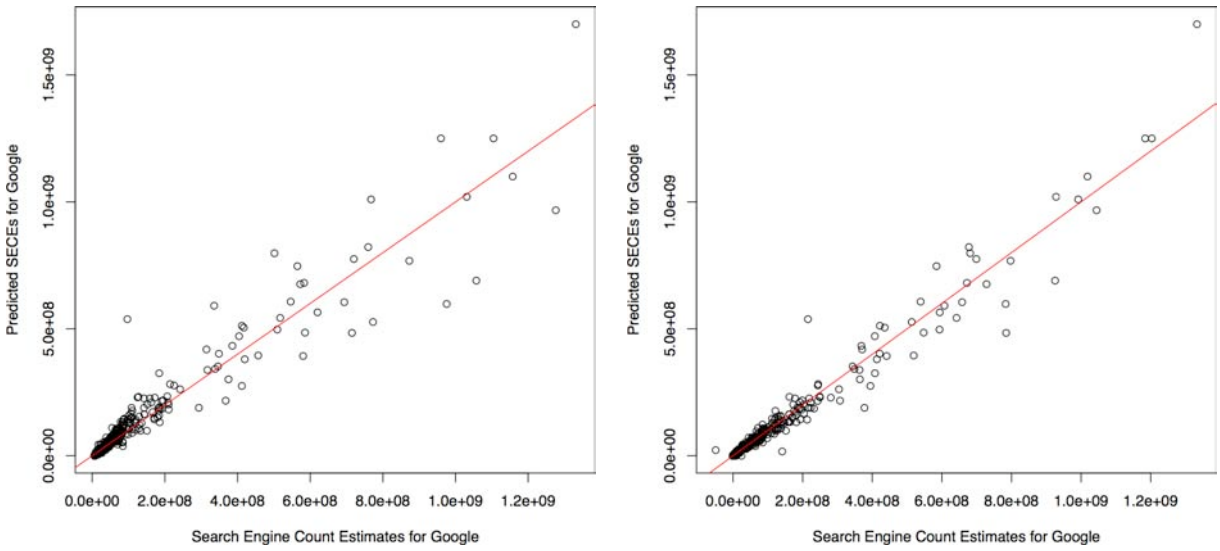


Figure 5. Study 5 – predicting Google SECEs via SECEs of other search engines (model performance, n=819 query terms).

Seen from a more general point of view it seems advisable to use both parallel-test and test-retest-reliability to control the data quality of SECEs. Since each measure casts a particular light on the reliability of SECEs both measures of reliability may be used in conjunction to study a phenomenon of interest from different vantage points. For instance, if a researcher wants to focus on highly volatile phenomena, e.g., expressed by concepts that are high on the agenda of global news, then it is the variation in time that is of central interest to the researcher, while test-retest reliability is expected to be low. In this case, however, one would like to know if volatile changes are also observed in other search engines. For this reason, the parallel-test-reliability should be considered.

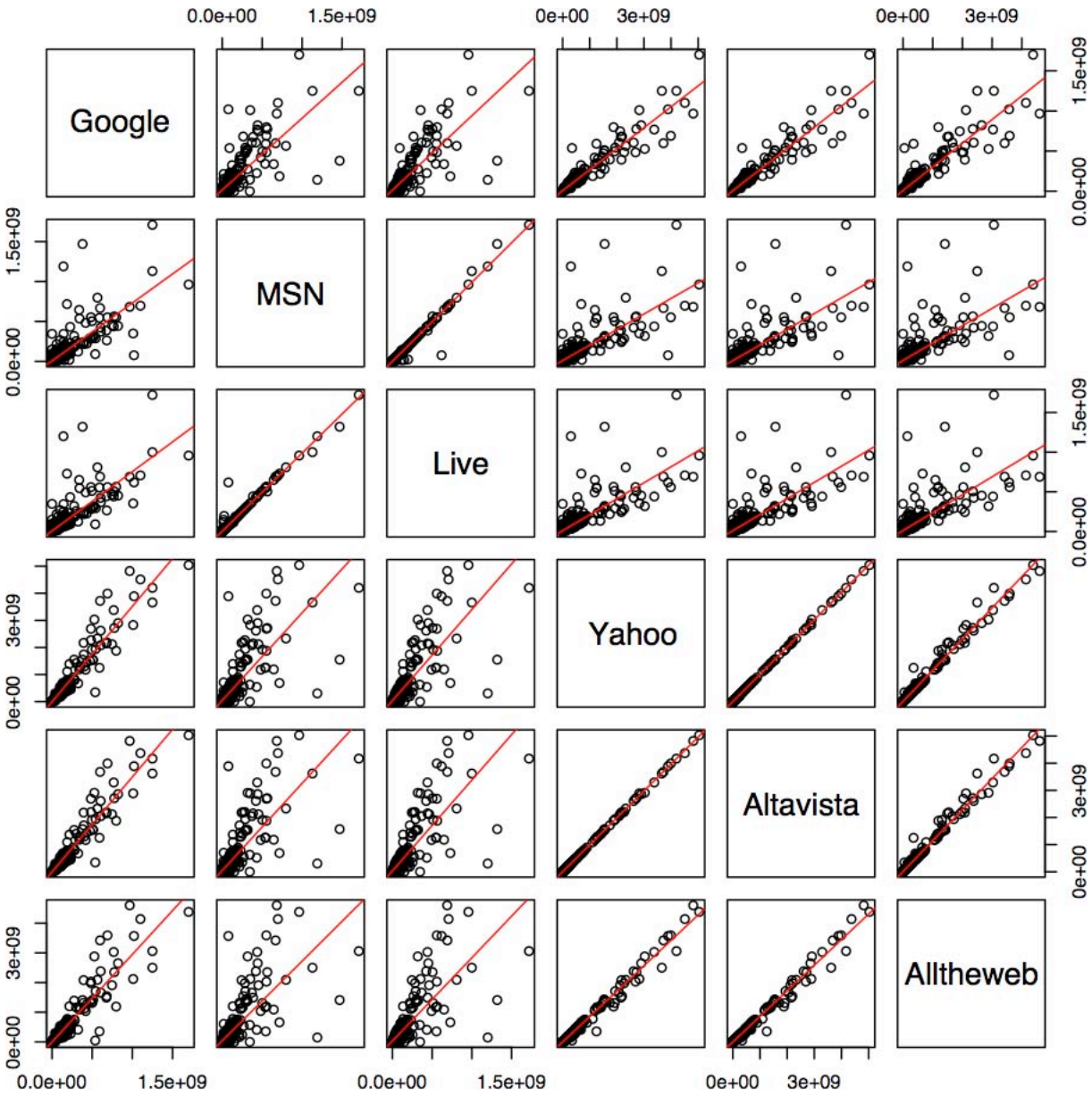


Figure 6. Study 5 – SECEs of Google, MSN, Live, Yahoo, Altavista, and Alltheweb in comparison (n = 819 query terms).

Study 6 – Validity

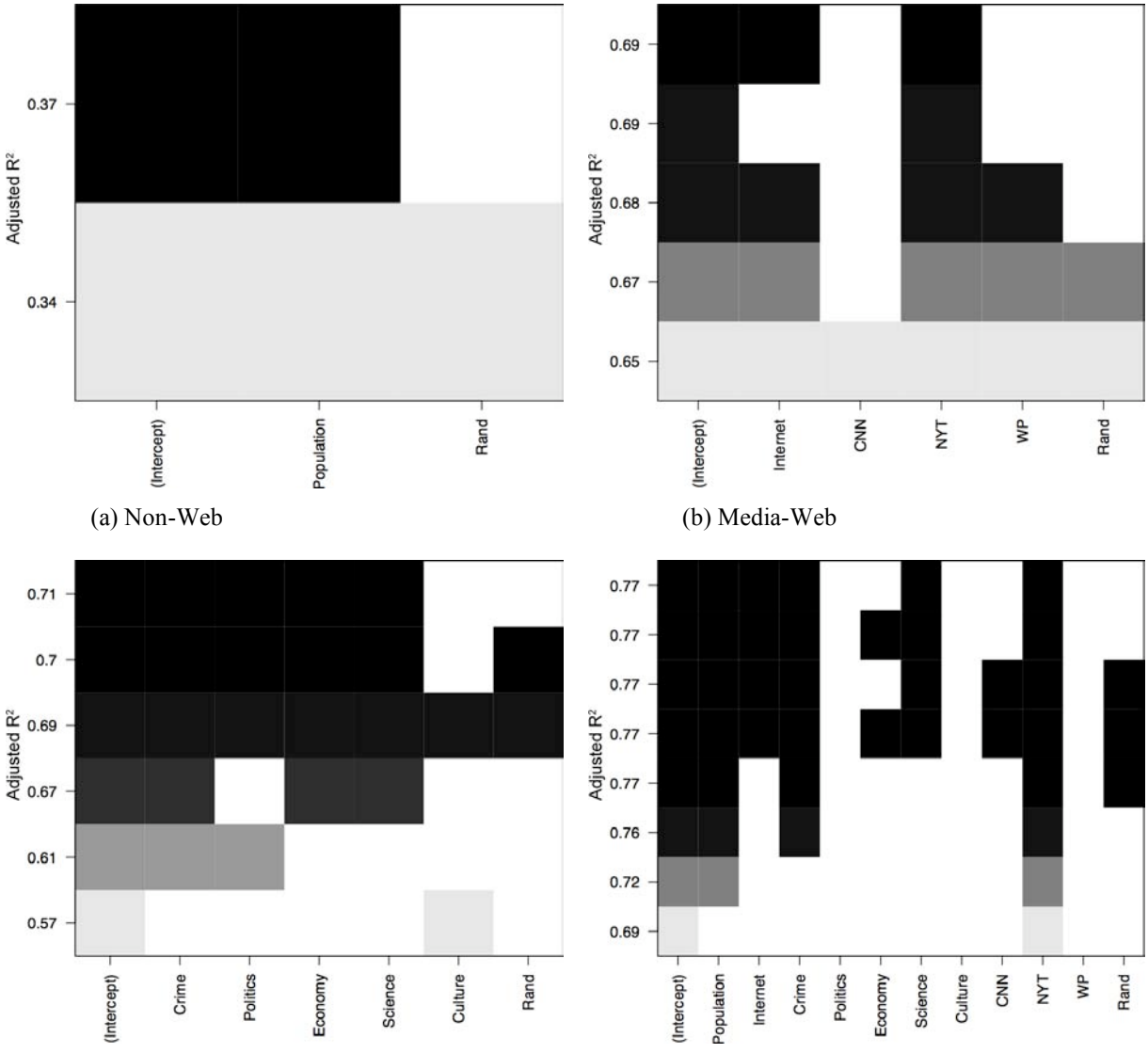
Validity of SECEs answers the question what SECEs mean. Ideally, validation studies are embedded in a more elaborate theory. Vice versa, examining the validity can promote theory development in a domain of interest by specifying the theoretical concepts, suitable empirical indicators, and their relationships.

In what follows, a pilot-study is presented that examined the construct validity of SECEs. Estimation of construct validity of SECEs was conducted via explorative model generation followed by model selection (Lahiri, 2001), which in turn was based on all-subset regression (Miller, 2002). Typical tasks of construct validation like construct explication (cf. Shadish, Cook, & Campbell, 2002, p. 64) and the estimation of convergent and discriminant validity were carried out within this framework. The construct chosen to be examined was the population size of 25 American cities with a population of more than 100,000 inhabitants, as evidenced by the magnitude of their SECEs.

Method

Again, the text-based browser Lynx along with a Perl script was employed to carry out all search operations and to process the results. The operating system used was Intel Mac OS X 10.5. The procedure used for construct

validation of SECEs had much in common with the procedure pursued in Study 5. In that study, specification of a model space was not necessary since SECEs of other search engines were used as predictors. By contrast, in Study 6 predictors to be considered were unknown since no theory seems to be available that explains if and to what extent population size of a city covaries with media coverage and/or with particular themes or topics (e.g., crime rate, science, economy, or culture). Therefore, in Study 6 the task of specifying a model space derived from the predictors used was imperative. A five-step procedure for selecting and examining predictor variables was administered to explain the dependent variable and thus to examine its construct validity. To initiate the search and quantification of predictor variables the first step was *model space specification*. Next, *model selection* was carried out on the basis of an all-subsets regression analysis (Miller, 2002). *Model performance* was established via multiple regression analyses (Figure 7), tests of *regression diagnostics* were conducted, and finally *convergent and discriminant validity* was investigated.



(a) Non-Web (b) Media-Web

(c) Topic-Web (d) Mixed

Figure 7. Study 6 – selection in a space of regression models following four different strategies ($n = 25$ query terms). Note the different scale resolutions on the y axes.

Model space specification. Construct explanation requires that the relationships between abstract concepts and their indicators have to be made transparent. Only then it is possible to trace back a score, e.g., a SECE, to conditions or variables that contribute to its magnitude. Tracing back SECEs for cities is not trivial since cities can have a large population for a number of reasons, and no theory seems to be available that can be deployed to select among them. For this reason, candidate predictor variables that could explain the population size of cities had to be specified in an explorative way. The approach for examining validity pursued in this paper was to compare and contrast four model generation strategies followed by model selection via regression analysis. Though this rationale is explorative it allows for testing competing models, which in turn is used to examine

convergent and discriminant validity. Each of the model generation strategies lead to a family of regression models that were analyzed in more detail in the subsequent steps.

(a) Non-Web. SECEs of the phenomenon of interest were analyzed by using data different from SECEs as predictor variables. In this study, population-size figures provided this kind of data. The source of the population-size figures for cities used in this study is the United States Census Bureau of 01.07.2006 (List of United States cities by population, 2008). The population sizes of the cities considered can be found in Table 5.

(b) Media-Web. SECEs of the phenomenon of interest, e.g., city, were analyzed by using SECEs as predictors obtained via Boolean conjunction of cities and media. An example of the search engine request that implements the media-Web strategy is *Boston AND CNN*. If this query leads to a high SECE then we may safely conclude that it is a good predictor for SECEs of Boston.

(c) Topic-Web. SECEs of the phenomenon of interest, e.g., city, were analyzed by using SECEs as predictors obtained via Boolean conjunction of cities and topics considered to be of importance. An example of the search engine request that implements the media-topic strategy is *Boston AND crime*. If this query leads to a high SECE then we may safely conclude that it is a good predictor for SECEs of Boston.

(d) Mixed. SECEs of the phenomenon of interest, e.g., city, were analyzed by jointly using the best predictors of all groups.

Table 5
Study 6 – Population of Selected United States Cities used to study the Validity of SECEs

City	Population
Chicago	2833321
Houston	2144491
Philadelphia	1448394
San Antonio	1296682
San Diego	1256951
Dallas	1232940
San Jose	929936
Detroit	918849
Jacksonville	794555
Indianapolis	785597
San Francisco	744041
Austin	709893
Memphis	670902
Fort Worth	653320
Baltimore	640961
Milwaukee	602782
Boston	590763
Seattle	582454
Denver	566974
Loisville	554496
Las Vegas	552539
Nashville	552120
Oklahoma City	537734

Data collection and model space specification were carried out as follows: Search was conducted manually by using Firefox 2.0.04. SECEs obtained for the name of a city, e.g., Boston, in Google were used as the dependent variable. The majority of the predictor variables (except population figures for cities) were SECEs obtained via Boolean search (e.g., Boston AND Internet, Boston AND crime, Boston AND science etc.). The predictor variables were derived from four classes of models

1. the population of the city considered (*non-Web*),
2. SECEs obtained on the basis of the Boolean queries Internet AND city, CNN AND city, "Washington Post" AND city (*media-Web*),
3. SECEs obtained on the basis of the Boolean queries crime AND city, politics AND city, economy AND city, science AND city (*topic-Web*), and
4. all predictors used in the previous analyses, i.e., population, Internet AND city (*mixed*).

Model Selection. Model assessment was carried out via an exhaustive search in the space of all regression models using all-subsets regression (Miller, 2002). As in Study 5, a random number predictor (*Rand*, ε) was included with each regression model. The outcome was a quantification of the goodness of all models across all four groups of model families considered (*non-Web*, *media-Web*, *topic-Web* and *mixed*). The model selection statistic used was adjusted R^2 (Lahiri, 2001).

Model Performance. A more fine grained assessment of model performance was carried out via multiple regression by way of predicting SECEs of cities on the basis of models identified via all-subsets regression (Miller, 2002).

Regression Diagnostics. To test whether assumptions of regression analysis were met the homoscedasticity, collinearity (correlations between predictor variables), correlation and normality of the residuals examined (Gross, 2003, Chap. 6).

Results

Model Space Specification and Model Selection. To identify and quantify models that predict the number of SECEs of cities an exhaustive search in the space of all regression models was carried out via all-subsets regression (Miller, 2002). The analysis was guided by four strategies for setting up regression models (non-Web, media-Web, topic-Web and mixed) outlined above. Results of this analysis are presented in Figure 7. In each of the four subfigures of Figure 7, single regression models are plotted horizontally while the corresponding model fit (adjusted R^2) is plotted on the ordinate. Therefore, models are ordered according to their model fit. Subfigure 8a shows the performance of two models. One model (black area) includes the intercept, population size and Rand (random numbers, ε) (adjusted $R^2 = .34$). The other model (grey area) includes the intercept and population size as predictors (adjusted $R^2 = .37$).

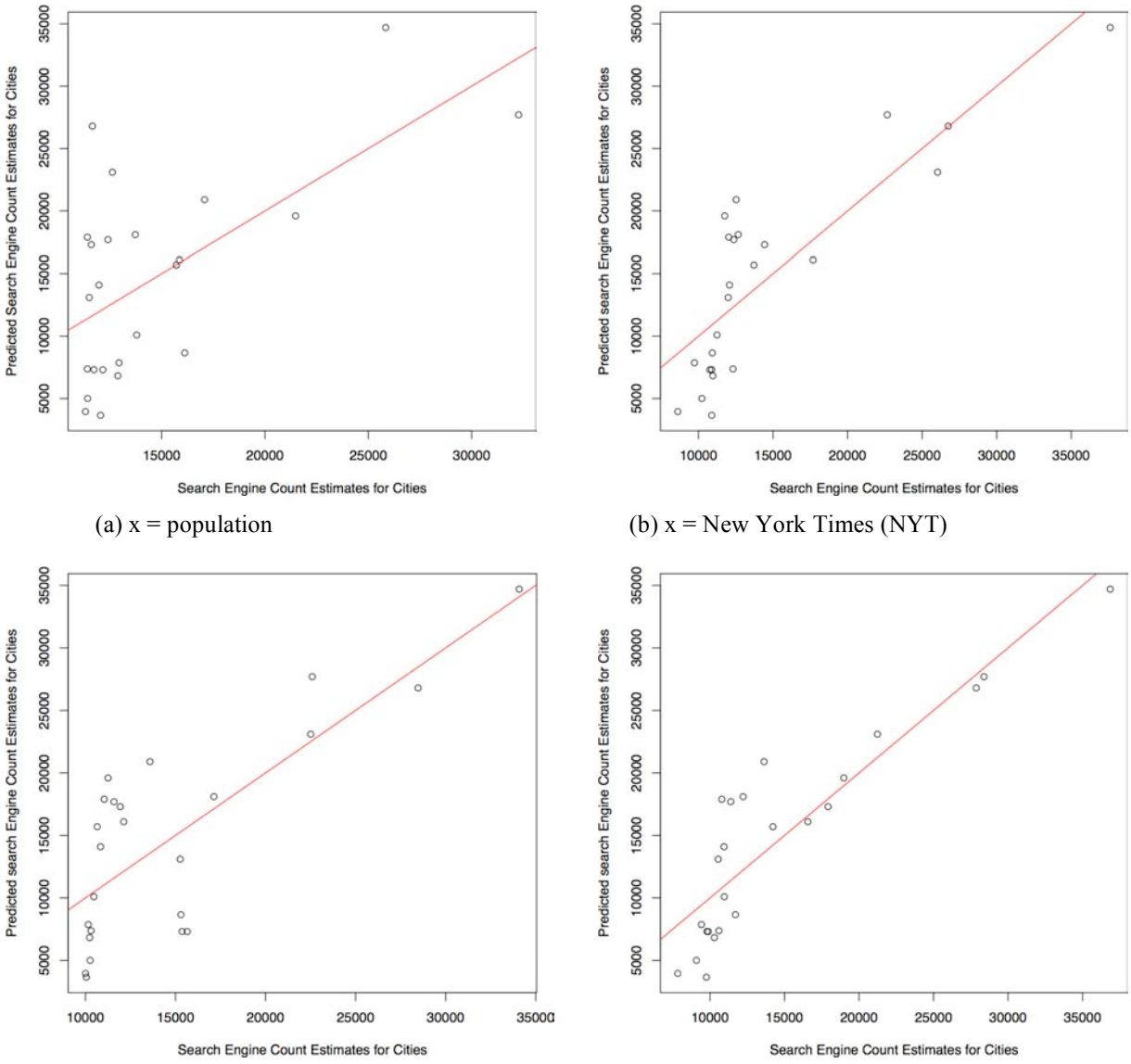
Model Performance. Multiple regression analysis was used to predict SECEs of cities on the basis of models identified by all-subsets regression (Miller, 2002). The results of this analysis are illustrated in Figure 7. Using cities as input, each subfigure plots observed (x-axis) against predicted SECEs and illustrates how well prediction of SECEs is possible. Subfigure 8a shows the effect of city population as a predictor. In subfigure 8b the effect of different media (Internet, CNN, New York Times, Washington Post) are presented. Subfigure 8c gives an outline of the effects of particular topics (Crime, Politics, Science, Culture). The effect of models set up on the basis of all predictor variables considered (Population, Crime, Politics, Science, Culture, Internet, CNN, New York Times, and Washington Post) are to be found in subfigure 7d.

The high potential of media variables as evidenced in subfigure 8b to predict SECEs comes as no surprise because we may assume that media repercussions lead to high SECEs. But media differ with regard to the degree they reflect influences that have a significant effect on the magnitude of SECEs. Study 6 shows that among all media types considered in this study the variable *New York Times* seemed to represent those influences best. The failure to statistically detect an influence of the variable *Politics* in the mixed model does not mean that it had no influence on the SECEs of cities. In the topic-Web model, the influence of the variable *Politics* is clearly given (subfigure 8c). In the mixed model, however, the media variables included in the model do the job of explaining the variable *Politics* (subfigure 7d) showing that it appears to have no direct influence on the SECEs of cities. To a lesser degree the same line of reasoning applies to other topic variables like *Crime*, *Culture*, *Economy*, and *Science*. Similar to the variable *Politics* their influence on SECEs on cities only surfaced when analyzed without taking media variables into account.

Regression Diagnostics. Least square estimation in regression analysis relies on homogeneous variance of the residuals. Its violation is usually called heteroscedasticity. Test for heteroscedasticity was carried out (Breusch-Pagan) but revealed no evidence for heterogeneous variance. Correlations between errors turned out to be very low and did not reach the level of significance. Most of the models set up and examined in this study were one-predictor models where collinearity is by definition not an issue. The only multi predictor model in this study was the best model for predicting SECEs for cities (Figure 7d). To test for collinearity of this model, variance inflation factors (VIF) were calculated. VIF factor scores for all predictors were low (all VIF < 2) indicating that predictor variables are not confounded.

Convergent and Discriminant Validity. The results of model performance and regression diagnostics provided evidence for convergent and discriminant validity. The expectation that the predictor variable considered contributes to an explanation of the dependent variable (population size of cities) by selecting variables associated with a target variable was confirmed. Almost all predictor variables covaried moderately or even strongly with the dependent variables as evidenced by moderate or high scores for adjusted R^2 across most models considered (cf. Figure 7). The negative result obtained for the collinearity tests of the full regression

model that included all predictor variables can be seen as evidence for discriminant validity. For instance, it cannot be expected that cities that are popular in the realm of science are also popular in the field of politics. Likewise, the population size of a city does not necessarily covary with the crime reported about a city via SECEs. Taken together, the evidence obtained for the convergent and discriminant validity contributes to a more differentiated understanding of what the magnitude of SECEs obtained for cities means.



(a) x = population

(b) x = New York Times (NYT)

(c) x = culture

(d) x = population, crime, and NYT

Figure 8. Study 6 – predicting Google SECEs for *cities* (model performance, $n = 25$ query terms).

Discussion

Study 6 exemplified a general heuristics for an evidence-based approach to examine construct validity.

Using the population size of cities to be predicted by the number of associated SECEs as a test case, it demonstrates the validation of SECEs by narrowing down and uncovering predictor variables that contribute to their magnitude. The model fit achieved was surprisingly good given that no domain theory could be used to select features required to design explanatory models. Instead, the result was achieved by using a explorative, but generalisable approach that facilitates testing of competing models. The method chosen rests on aggregated accounts, which in turn are based on a set of concepts (cities) and have to be interpreted accordingly. This means, the validity results for a particular city may or may not match closely with the aggregated results. Validating SECEs of cities is a typical task, because one can assume that the magnitude of SECEs of cities is influenced by many variables. This approach can also be used to work towards a theory in a domain of interest by highlighting the role of more general concepts and their empirical indicators in a nomological network of relationships.

Conclusion

Search engines do not only search, they also elicit and organize huge amounts of data. A large proportion of these data is provided in a qualitative way, i.e., references providing pointers to documents. It is via search engine count estimates (SECEs) that many search engines express aspects of qualitative data in a quantitative way. The results of the studies presented in this paper found evidence of a good data quality of search engine estimates in terms of objectivity, reliability, and – to a lesser degree – the validity of SECEs. Clearly, the size and organization of indices of search engines change dynamically and data gleaned from search engines is not free of noise. An overreliance on a particular search engine needs to be avoided since the technologies, e.g., the algorithms, are subject to change. A search engine company may change or lose its leading position. But these problems do not mean that the data quality renders SECEs useless. Collecting and using search count estimates that is paralleled by meticulous quality control helps to identify noise in the data. Methods of statistical meta-analysis (Hunter & Schmidt, 2004; Hartung et al., 2008) facilitate the mutual quality control and aggregation of data gleaned from different search engines.

Previous Objections

How do the results obtained relate to the objections leveled against SECEs? Users expect search engines to provide URLs associated with a search query to satisfy their information needs. They also expect this service to be delivered in an extremely short time. Search engines are primarily designed to fulfill these expectations. Presenting count estimates of documents may work as additional information. The credibility and thus the commercial success of search engines would be at risk if SECEs were implausible or counterintuitive. In short, seen from the viewpoint of search engine companies, count estimates do seem to matter, but at the moment they do not seem to play a key role. This ambivalence is echoed in the way SECEs are received by many researchers. While there is a burgeoning trend to make use of them (Cilibrasi & Vitanyi, 2007; Gligorov et al., 2007) several authors have raised concerns about the data quality of SECEs (Bar-Ilan, 1999; Rousseau, 1999; Wouters et al., 2004). Following is a discussion of some issues raised around SECEs in light of the studies of this paper.

Just estimates. Search engines do not report on the exact number of documents associated with a search query. Instead, an estimation procedure is applied that returns rounded figures. This becomes obvious when search queries return count estimates of high-frequency terms, e.g., *Internet*. The fact that SECEs are estimates does not in itself mean that SECEs are of poor data quality. For instance, the results of Studies 4 and 5 attest that these estimates tend to be of good reliability. Poor test-retest reliability was predominantly found among extremely rare concepts, e.g., *ahgareseh*. Note that almost none of the concepts used to study test-retest reliability was part of the language used to describe the events that were high on the agenda of the global news in the time of data recording (e.g., *olympic, Georgia, credit, crisis*). Clearly, the usage of such concepts would have resulted in higher variability and thus lower test-retest reliability. Still, it would have been interesting to focus just on concepts currently in the global news. Here, a low test-retest reliability, but a high parallel-test reliability can be expected.

Deep Web. Most search engines use crawlers to create and update their index on a regular basis. A crawler is a piece of software that follows links on Web pages thereby finding other pages. This procedure works well with regard to static pages that are linked to other pages, but it fails with regard to pages that are dynamic or with not links to other pages. The summary term for pages that can not be found by search engines is *deep Web* (Bergman, 2001). The existence of the deep Web may threaten the validity of SECEs if this means that a number of content areas is systematically excluded from the index of search engines. Study 6 found satisfying results on the validity of the concept *city*. But clearly, this study was of a pilot character and more studies on the validity of SECEs are needed. It is conceivable that there are SECEs of some groups of concepts that can be validated well, while others lead only to poor validation scores. Whether the deep Web may or may not influence the validity of SECEs of some groups of concepts remains an open issue.

Boolean Operators. There are indications that Boolean queries lead to anomalous SECEs (Liberman, 2005). In fact, Study 2 found distorted results for all Boolean queries (tested via Google) except for conjunctions of type *a* AND *b*. This type of conjunction was found free of distortions and it was used extensively in Study 6.

Small Result Overlap Between Search Engines. Some studies report on low result overlap of search engines (Spink et al., 2006; dogpile.com, 2007). However, the far-reaching conclusions in the studies mentioned did not match up with the method actually used. The small result overlap was found just on basis of the first result page returned. In contrast, Study 5 of this paper found a high result overlap and a high parallel-test reliability on the basis of all concepts used.

Potential Applications

What are potential applications that come into reach on the basis of a statistical analysis of search engine estimates? Search engines reflect aspects of reality, or to be more precise, they reflect the sphere of documents available on the Internet. Vice versa, the statistical analysis of search engine estimates facilitates measurements of phenomena related to this sphere. The results outlined in the present paper can be used to work towards an *Internet resonance diagnostics*, i.e., a method to measure the magnitude of global or local events. The repercussions of natural catastrophes, the outcome of elections in a particular country, the beginning or end of a war, or the global echo of a company's marketing campaign are all examples of phenomena the repercussions of which could be measured on the basis of SECEs.

Future Work

A number of tasks has to be accomplished before Internet resonance diagnostics can be carried out. Firstly, while the work outlined in this paper is a first contribution to a rigorous examination of SECEs more studies are required to examine their objectivity, reliability, and validity. In particular, consideration of both test-retest reliability and parallel-test reliability holds the key for a sound analysis of SECEs. Secondly, studies both content-related and methodologically oriented are needed to examine phenomena of change (e.g., Internet resonance on a war, global warming, or the introduction of a new and highly attractive computer game). Measurement of change that is based on SECEs needs to disentangle effects of the changing phenomenon itself from noise that may be caused by the overall setup of search engines (e.g., change of index size or index organisation). Thirdly, there is a need for calibration studies to find out context-dependency of the magnitude of SECEs. It may well be that events lead to SECEs of a higher or smaller magnitude just because of their domain of origin (e.g., politics, culture). Fourthly, work towards a general measure of Internet resonance is needed. Ideally, the analysis of SECEs should be independent of a particular search engine and robust with regard to internal re-organisation or re-indexing of search engines (Bar-Ilan, 1999). The results of Study 5 in this paper indicate that this objective can be achieved, since correlations between SECEs of different search engines turned out to be high. Finally, a thorough study of different event types is required. For instance, measurement sensitivity for sudden global events (e.g., the 9-11 attacks) may differ from global events with slowly increasing or with volatile intensity of media coverage (e.g., the credit crunch crisis that started 2008).

Search engines are collecting data on all published aspects of modern societies. In line with previous observations this study found that Boolean queries (except conjunctions) lead to anomalous numbers of SECEs. In general, however, the results of this work indicate that the quality of SECEs in terms of objectivity, reliability, and validity is surprisingly good. If elicitation and analysis of SECEs is guided by a thorough methodology, we may expect that data gleaned from search engines will cast light on a variety of phenomena from politics, sociology, economics, linguistics, and psychology.

Acknowledgments

The author would like to thank Nigel Fielding, Katherine Lacey, Andrew Page, Stephan Weibelzahl and the anonymous reviewers of the International Journal of Internet Science for insightful comments on previous versions of this paper.

References

Anonymous. (2005). *Google hits extraction*. Retrieved December 16, 2008, from http://www.cantonese.sheik.co.uk/phorum/read.php?2,42057

Bagrow, J. P., & ben-Avraham, D. (2005). On the Google-fame of scientists and other populations. *Proceedings of the American Institute of Physics Conference, 779*(1), 81–89. Retrieved December 16, 2008, from http://arxiv.org/abs/physics/0504034

Bar-Ilan, J. (1999). Search engine results over time: A case study on search engine stability. *Cybermetrics, 2/3*(1). Retrieved December 17, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

Bar-Ilan, J. (2001). Data collection on the Web for informetric purposes: A review and analysis. *Scientometrics, 50*(1), 7–32.

Bar-Yossef, Z., & Gurevich, M. (2006). Random sampling from a search engine's index. In *WWW '06: Proceedings of the 15th international Conference on World Wide Web* (pp. 367–376). New York: ACM Press.

Bergman, M. K. (2001). The deep Web: Surfacing hidden value. *Journal of Electronic Publishing, 7*(1). Retrieved December 16, 2008, from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. In *WWW7: Proceedings of the seventh International Conference on World Wide Web 7* (pp. 379–388). Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Britannica. (2008). *Index terms of the online-version of Encyclopædia Britannica*. Retrieved December 16, 2008, from http://www.britannica.com/bps/browse/alpha/a

Carmines, E. G., & Zeller, R. A. (1991). *Reliability and validity assessment.* Thousand Oaks, CA: Sage.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transaction on Knowledge and Data Engineering, 19*(3), 370–383.

dogpile.com. (2007). *Different search engines – different results: A study by dogpile.com*. Retrieved December 16, 2008, from http://www.infospaceinc.com/onlineprod/Overlap-DifferentEnginesDifferentResults.pdf

Flack, V. F., & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician, 7*(1), 84–86.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Gligorov, R., Kate, W., Aleksovski, Z., & Harmelen, F. van. (2007). Using Google distance to weight approximate ontology matches. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web* (pp. 767–776). New York: ACM.

Google, Inc. (2008). *Developer resources*. Retrieved March 8th, 2008, from http://code.google.com

Gross, J. (2003). Linear regression. *Lecture Notes in Statistics 175*. Berlin: Springer.

Gruijter, D. N. M. D., & Kamp, L. J. T. v. d. (2007). *Statistical test theory for the Behavioral Sciences*. Boca Raton, FL: Chapman & Hall/CRC.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3,* 1157–1182.

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical meta-analysis with applications.* New York: Wiley.

Henzinger, M. R., Heydon, A., Mitzenmacher, M., & Najork, M. (1999). Measuring index quality using random walks on the Web. In *8th International World Wide Web Conference, WWW8,* May 1-5, 2001, Toronto, Canada.

Hundt, M., Biewer, C., & Nesselhauf, N. (Eds.). (2007). *Corpus linguistics and the Web*. Amsterdam: Rodopi.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Krug, M. (2006). *Modern methodologies and changing standards in english linguistics*. Annual meeting of the Spanish Association for English and American Studies (AEDEAN, Asociación Espanola De Estudio Anglo-Norteamericano), Spain.

Lahiri, P. (2001). *Model selection*. Monograph Series, Vol. 38. Beachwood, OH: Institute of Mathematical Statistics Lecture Notes.

Landers, B. (2008). *PyGoogle: A Python interface to the Google API*. Retrieved December 16, 2008, from http://pygoogle.sourceforge.net

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science, 280,* 98–100.

Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2005). The freshness of Web search engines' databases. *Journal of Information Science, 32*(2), 131–148.

Li, W. (1992). Random texts exhibit Zipf 's law-like word frequency distribution. IEEETIT: IEEE *Transactions on Information Theory, 38*. Retrieved December 16, 2008, from http://www-personal.umich.edu/ ~warrencp/WLi.pdf

Liberman, M. (2005). *Posting to language log on Boolean search*. Retrieved December 16, 2008, from http://158.130.17.5/~myl/languagelog/archives/001831.html

List of United States cities by population. (2008, December 16). In Wikipedia, The Free Encyclopedia. Retrieved December 16, 2008, from http://en.wikipedia.org/w/index.php?title=List_of_United_States_cities_by_ population&oldid=258384510

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54,* 217–224.

Mayr, P., & Tosques, F. (2005). *Google Web APIs: An instrument for webometric analyses?* Poster presented at the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI) in Stockholm, Sweden. Retrieved December 16, 2008, from http://bsd119.ib.hu-berlin.de/~ft/index e.html

Miller, A. (2002). *Subset Selection in Regression* (2nd ed.). London: Chapman & Hall/CRC.

Odom, L. R., & Morrow, J. R. J. (2006). What's this r? A correlational approach to explaining validity, reliability and ob jectivity coefficients. *Measurement in Physical Education and Exercise Science, 10*, 137–145.

Pullum, G. K. (2004). *Webhits on Google per gigapage: A replacement proposal.* Retrieved March 6, 2008, from http://itre.cis.upenn.edu/~myl/languagelog/archives/000958.html

Rousseau, R. (1999). Daily time series of common single word searches in Altavista and Northernlight. *Cybermetrics, 2/3*(1). Retrieved December 16, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/ v2i1p2.html

Rusmevichientong, P., Pennock, D. M., Lawrence, S., & Giles, C. L. (2001). Methods for sampling pages uniformly from the World Wide Web. In *AAAI Fal l Symposium on Using Uncertainty Within Computation* (pp. 121–128).

Schuster, D., & Schill, A. (2007). NL sampler: Random sampling of Web documents based on natural language with query hit estimation. *In Proceedings of the 2007 ACM symposium on Applied computing (SAC '07).*

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Miiffin.

Spink, A., Jansen, B. J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major Web search engines. *Information Processing and Management, 42*(5), 1379–1391.

W3schools. (2008). *Browser statistics*. Retrieved December 16, 2008, from http://www.w3schools.com/ browsers/browsers stats.asp

Whaley, C. P. (1978). Word nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17,* 143–154.

White, H. (1980). A heteroscedastic-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica, 48,* 817–838.

Wouters, P., Hellsten, I., & Leydesdorff, L. (2004). Internet time and the reliability of search engines. *First Monday, 9*(10). Retrieved December 16, 2008, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/ issue/view/176

Yahoo, Inc. (2008). *Yahoo developer network*. Retrieved December 16, 2008, from http://developer.yahoo.com/ search/

Zellner, A., Keuzenkamp, H. A., & McAleer, M. (Eds.). (2001). *Simplicity, inference and modeling: Keeping it sophisticatedly simple*. Cambridge, MA: Cambridge University Press.

Zipf, G. (1932). *Selected studies of the principle of relative frequency in language.* Cambridge, MA: MIT Press.

Appendix A

One-Tuple Terms Gleaned from Encyclopædia Britannica (n=819)

The following set of one-tuple terms was derived from the Online version of Encyclopædia Britannica (4-4-2008). The concepts were used as search terms in studies 1 and 5 and formed the basis for the low, medium, and high frequency concepts employed in studies 2–4. Note that publication of this article will change the magnitude of SECEs reported.

abbreviation, accountant, acrobatics, acrodynia, actor, actress, actuary, addiction, adolescence, adulthood, aeronautics, Afghanistan, Africa, agnosticism, agriculture, agroforestry, agronomy, ahgareseh, ahje, aircraft, airfield, airframe, airmail, airport, airscrew, Albania, alcoholism, ale, Algeria, alguacile, allergy, allspice, amateur, ammunition, anachronism, analvos, anarchism, andiron, Andorra, androgyny, anemochory, anemotaxis, Anglicanism, Angola, animism, anirmoksa, annuity, Antarctica, anthropocentrism, anthropology, antidepressant, antique, antiseptic, antiserum, antitoxin, antler, apothecary, appendicitis, apperception, appetizer, apprenticeship, aquarium, Arabization, arboretum, arbovirus, archaeology, archery, archetype, architecture, architrave, archives, archivolt, Arctic, Argentina, aristocracy, Aristotelianism, arithmetic, arithmomancy, Arithmometer, Armada, Armenia, army, arson, art, arteriosclerosis, artillery, artist, asceticism, Asia, asphalt, Assyriology, astrology, astronaut, astronomy, astrophysics, atheism, athletics, atomizer, attorney, audiocassette, Australia, Austria, authority, autohypnosis, autoimmunity, automation, automaton, automimicry, automobile, autoprotolysis, autopsy, autotomy, avacchedakata, aviary, aviation, aviculture, axiology, axiom, Azerbaijan, backhoe, backlighting, bacteria, Badme, Bahrain, bailee, Balkanization, Balkans, ballista, ballistics, ballistocardiogram, ballistocardiography, ballistospore, balustrade, banderilla, banderillero, Bangladesh, bankruptcy, Barbados, barber, barometer, Bartmannkrug, basketry, bathroom, batterie, battleship, beekeeping, beeswax, beetling, behaviour, Belarus, Belgium, belief, Belize, betatron, beverage, Bhutan, bibliography, bilingualism, biliprotein, billboard, bioarchaeology, bioavailability, biocentrism, bioceramics, biochronology, biocontrol, bioengineering, bioethics, bio-fuel, biogeography, biology, biomaterials, biometrics, bioplastics, biosphere, biotechnology, biotelemetry, bioterrorism, bioturbation, Blaberidae, blacklist, blasthole, blasting, blueprint, bobbinet, bodyguard, Bolivia, bookmobile, bookplate, bookseller, bookselling, borderlight, boredom, botany, botnet, Botswana, bottling, boxcar, boycott, brainwashing, Brazil, brazilwood, brazing, brewing, brickwork, broadcasting, Brunei, buccaneer, Bulgaria, bulking, bullfighter, bullfighting, bumblebee, bureaucracy, burgage, burladero, burnishing, Burundi, butcherbird, cabaret, cabinetmaking, caboose, cafeteria, calculator, calligraphy, Cambodia, Cameroon, camping, campus, Canada, capitalism, caries, carob, carpooling, cartoonist, carving, caryopsis, cassette, castration, cataclastite, catapult, catfish, catgut, catwalk, cave, caviar, celibacy, cellophane, cemetery, censorship, census, centipede, ceramics, Ceravix, cereal, certification, cesta, cestodiasis, Chad, chancellor, charadriiform, charcoal, chastity, chauvinism, cheating, chemiluminescence, chemistry, chessboard, chessmen, chicuelina, childhood, Chile, China, chivalry, Christianity, chronobiology, chronogram, chronology, chronon, chronophotography, cigar, cigarette, cigarillo, citizenship, city, civilization, cleaving, clepsydra, cleruchy, clinic, clinograde, clown, cobia, cobiron, cockpit, coeducation, coffin, cognition, coin, coinsurance, coir, coking, collectible, cologne, Colombia, commerce, comminution, Comoros, composting, compressor, computer, conduct, Confucianism, conscience, consciousness, container, contemplation, contest, contraception, cookbook, cooking, cordage, corncob, cornstarch, coroner, corporation, cowboy, cradleboard, crayon, creationism, creativity, criminology, Croatia, crosswind, crowding, cryoprotectant, cryopump, cryosurgery, cryptanalysis, cryptography, cryptology, crystallography, Cuba, cuisine, culture, cutlery, cutwork, cybercrime, cyberlaw, cybernetics, cycloalkane, Cyprus, cytotrophoblast, Dadaist, dagger, dairying, damask, dawn, deafness, death, decaffeination, decarboxylation, decarburization, decolonization, defeathering, deforestation, deglaciation, deinstitutionalization, demantoid, demilitarization, democracy, democratization, Denmark, dentistry, deodorization, desert, detergent, dialogue, dinnerware, diplomacy, discipline, disease, disinfectant, dissidence, divorce, dizziness, Djibouti, documentation, dogfighting, doll, domicile, Dominica, dough, drilling, drowning, duel, dunite, dust, dye, dynasty, earphone, earwax, echolocation, ecofeminism, ecology, ecomuseum, economics, ecosystem, ecoterrorism, ecotourism, Ecuador, education, eel, Egypt, eiderdown, electroceramics, electrochemiluminescence, electrochemistry, Electrofax, electroforming, electrogalvanizing, electrojet, electrometallurgy, electronics, electrophysiology, electroplating, electropolishing, electropositivity, electroreception, electrotyping, elephant, embalming, embroidery, embryology, enology, enterotoxin, entertainment, Epicureanism, epigenetics, epizoochory, eponym, eraser, Eritrea, Estonia, Ethiopia, ethnobotany, ethnography, ethnohistory, ethnolinguistics, ethnomusicology, ethnopharmacology, ethnopsychiatry, ethology, etymology, eulogy, eunuch, Eurasia, Eurocommunism, Eurocurrency, Eurodollar, Eurogroup, Europe, exhibition, exploration, exsanguination, extirpation, Exubera, factory, faena, falconry, famine, fat, fecundity, feeblemindedness, feminism, ferryboat, fibreglass, Finland, fireboat, fireproofing, fisherman, fishery, flame, flatware, flexography, flood, flowchart, food, footbridge, forestry, fortnight, foundry, fountain, foxhunting, France, freight, freshwater, friendship, frostbite, fume, fumigation, funding, furnace, furniture, Gabon, gagging, Galvus, gambling, gaonera, garbage, garnishment, gastronomy, gazetteer, gelatinization, gematria, gemmulation,

gemmule, genetics, genocide, genomics, geochemistry, geography, geometry, geomorphology, geophysics, geopolitics, Georgia, geotectonics, gerenuk, Gereshk, Germany, Ghana, gibberfish, giftbook, gigantism, gingivitis, glacier, glaciology, glass, glassblowing, glassware, glossary, glossography, gnawing, goblet, godspell, goosefish, gouache, government, grammar, grandparent, grassland, gravel, grease, Greece, greenhouse, Greenland, greeting, Grenada, grief, grindability, guardian, Guatemala, Guinea, gulf, gunport, Guyana, gymnastics, gyroscope, hairdresser, hairdressing, hairyfish, Haiti, Halloween, handcuffs, handicraft, handwriting, hanging, happiness, harbour, hardboard, hardstone, harem, hat, hatchetfish, headache, headlight, headphone, headscarf, headwear, health, heartburn, heaven, hedge, hemoglobinopathy, hennin, hepatitis, heptane, heraldry, herbicide, herbivore, herd-book, heresy, heterodonty, hexabromocyclododecane, hieroglyph, Hiranyagarbha, Hirta, histochemistry, historiography, history, homosexuality, Honduras, horsemanship, horseshoe, horticulture, hospital, hotbed, hotel, houseboat, household, houseware, housing, hubris, hulling, humanism, humanitarianism, humidifier, Hungary, hydrobiology, hydrofining, hydrofluorocarbon, hydrofoil, hygiene, hymnbook, hypervalence, hyphen, Iceland, iconodule, idea, ideogram, ideography, idolatry, imagination, immigration, immunotherapy, incense, incest, incineration, incinerator, incisor, India, indoctrination, Indonesia, Paraguay, puzzle, pyranose, Pyrex, suburb, subway, Tajikistan, tampon, Tanzania, tapestry, tarnish, tasting, tatami, taxation, taxidermy, teaching, teakettle, teapot, technology, telecommunication, telegraph, telephone, telephotography, teratology, terrorism, textile, Thailand, thawing, theology, thermocouple, thermodynamics, thimble, thought, tile, tintinnid, tinware, tire, toadstool, tobogganing, Togo, tokonoma, toll, tomography, tonadilla, Tonga, tonneler, tonometer, tonos, tool, tooling, topazolite, topiary, topology, toponomastics, toponymy, torch, torpor, torture, totem, toupee, tourism, town, towpath, toxicology, toxoid, toy, transesterification, transponder, transsexualism, transumpt, transvestism, trapezohedron, travel, trepanning, trigonocephaly, trigonometry, trihexaflexagon, trincherazo, trolling, Trotskyism, trunnion, tugboat, Tunisia, turbine, turboramjet, tureen, Turkey, Turkmenistan, Tuvalu, twin, typewriter, typography, typology, Uganda, Ukraine, underwriting, unemployment, uniform, Unitarian, university, upholstery, urbanization, Uruguay, uvarovite, Uzbekistan, vacation, Vanuatu, vegan, vegetarian, Venezuela, victimology, videocassette, Vietnam, vigilante, virginity, visbreaking, wallowing, wallpaper, war, warehouse, warehouseman, warhead, warship, wasp, water, watermelon, waterpower, wax, waxplant, waxwing, weather, weatherfish, wetland, Wewoka, whaling, wheel, whirlpool, whitefish, wholesaling, widowhood, wig, windmill, wine, wire, wiretapping, witchcraft, women, woodcarving, woodland, writing, xylography, yarn, Yemen, youth, Zambia, zeitgeber, Zimbabwe, zipper, Zipporah, zoo, zoogeography, zoology, zooplankton, Zostavax, Zwinglian

Appendix B

Table B1

Study 4 on Test-Retest Reliability of SECEs – Product Moment Correlation Scores of Concepts (n = 15) Generating a High Number of SECEs. Calculations based on Intervals of 10 days (day 1 – day 161)

Day	1	11	21	31	41	51	61	71	81	91	101	111	121	131	141	151
11	.99***															
21	.95***	.97***														
31	.97***	.98***	1.00***													
41	.83***	.78***	.64**	.68**												
51	.96***	.98***	1.00***	1.00***	.69**											
61	-.11	-.12	-.08	-.05	-.35	-.10										
71	.64**	.70**	.67**	.67**	.35	.65**	-.01									
81	.94***	.97***	1.00***	.99***	.63**	1.00***	-.08	.66**								
91	.90***	.94***	.99***	.98***	.53*	.98***	-.03	.68**	.99***							
101	.88***	.92***	.94***	.94***	.51*	.93***	.05	.85***	.94***	.95***						
111	.61*	.68**	.82***	.78***	.11	.78***	.15	.59*	.83***	.89***	.83***					
121	.85***	.91***	.95***	.94***	.48	.93***	-.06	.80***	.95***	.97***	.97***	.89***				
131	.63**	.70**	.84***	.80***	.16	.80***	.09	.56*	.85***	.90***	.83***	.99***	.89***			
141	.34	.44	.60*	.54*	-.12	.56*	-.12	.48	.62**	.70**	.64**	.89***	.75***	.89***		
151	.06	-.02	-.21	-.15	.51*	-.17	.09	-.19	-.24	-.33	-.27	-.62**	-.36	-.59*	-.82***	
161	.75***	.80***	.86***	.83***	.44	.85***	-.37	.68**	.87***	.87***	.82***	.76***	.86***	.78***	.72**	-.49*

*p < .05. **p < .01. ***p < .001.

Table B2

Study 4 on Test-Retest Reliability of SECEs – Product Moment Correlation Scores of Concepts (n = 15) Generating a Medium Number of SECEs. Calculations based on Intervals of 10 days (day 001 – day 161)

Day	1	11	21	31	41	51	61	71	81	91	101	111	121	131	141	151
11	1.00***															
21	1.00***	1.00***														
31	1.00***	1.00***	1.00***													
41	.98***	.98***	.98***	.98***												
51	1.00***	1.00***	1.00***	1.00***	.98***											
61	.98***	.97***	.97***	.97***	.97***	.97***										
71	.99***	1.00***	.99***	1.00***	.99***	1.00***	.98***									
81	.99***	.99***	.99***	.99***	1.00***	.99***	.98***	1.00***								
91	.99***	.99***	.99***	.99***	1.00***	.99***	.98***	1.00***	1.00***							
101	.96***	.96***	.95***	.95***	.99***	.95***	.96***	.97***	.98***	.99***						
111	-.49*	-.50*	-.51*	-.51*	-.38	-.53*	-.45	-.48*	-.43	-.40	-.29					
121	-.09	-.09	-.11	-.12	-.01	-.14	-.04	-.10	-.05	-.02	.07	.85***				
131	-.84***	-.84***	-.85***	-.85***	-.79***	-.86***	-.83***	-.84***	-.81***	-.79***	-.72**	.78***	.53*			
141	-.67**	-.66**	-.66**	-.65**	-.64**	-.65**	-.73***	-.64**	-.63**	-.63**	-.59*	.35	-.04	.67**		
151	-.95***	-.95***	-.95***	-.95***	-.93***	-.95***	-.90***	-.94***	-.95***	-.95***	-.91***	.41	.03	.75***	.63**	
161	-.95***	-.94***	-.94***	-.94***	-.89***	-.94***	-.91***	-.92***	-.91***	-.91***	-.86***	.50*	.08	.81***	.75***	.97***

*p < .05. **p < .01. ***p < .001.

Table B3

Study 4 on Test-Retest Reliability of SECEs – Product Moment Correlation Scores of Concepts (n = 15) Generating a Low Number of SECEs. Calculations based on Intervals of 10 days (day 001 – day 161)

Day	1	11	21	31	41	51	61	71	81	91	101	111	121	131	141	151
11	.92***															
21	.77***	.96***														
31	.74***	.94***	1.00***													
41	.67**	.90***	.98***	.99***												
51	.56*	.82***	.92***	.94***	.98***											
61	.34	.62**	.76***	.79***	.88***	.95***										
71	.26	.53*	.68**	.72**	.82***	.91***	.99***									
81	.10	.36	.51*	.55*	.68**	.80***	.95***	.98***								
91	.21	.47	.61**	.65**	.76***	.87***	.98***	.99***	.99***							
101	.73***	.93***	.98***	.99***	.99***	.96***	.84***	.78***	.63**	.72**						
111	.95***	.93***	.83***	.81***	.77***	.70**	.52*	.45	.32	.42	.84***					
121	.41	.04	-.23	-.27	-.33	-.39	-.46	-.47	-.44	-.42	-.22	.34				
131	-.05	-.42	-.66**	-.69**	-.73***	-.76***	-.74***	-.72**	-.62**	-.64**	-.64**	-.13	.89***			
141	-.18	-.55*	-.76***	-.79***	-.82***	-.84***	-.80***	-.77***	-.65**	-.69**	-.75***	-.28	.81***	.99***		
151	-.27	-.62**	-.81***	-.84***	-.87***	-.88***	-.82***	-.78***	-.65**	-.70**	-.80***	-.37	.75***	.97***	1.00***	
161	-.34	-.67**	-.85***	-.87***	-.89***	-.90***	-.83***	-.78***	-.64**	-.70**	-.83***	-.42	.71**	.95***	.99***	1.00***

*p < .05. **p < .01. ***p < .001.