Language, culture and traversing the scholarly evaluation landscape.

Zuccala, Alesia Ann

Published in:
The Evaluation of Research in the Social Sciences and Humanities. Lessons from the Italian Experience.

Publication date:
2018

Document version
Peer reviewed version

Citation for published version (APA):
Zuccala, A. A. (2018). Language, culture and traversing the scholarly evaluation landscape. In A. Bonaccorsi (Ed.), *The Evaluation of Research in the Social Sciences and Humanities. Lessons from the Italian Experience.* (pp. 395-411). Switzerland: Springer.

Author's preprint version, August 10, 2017.

Language, Culture and Traversing the Scholarly Evaluation Landscape

Alesia Zuccala
Email a.zuccala@hum.ku.dk
Department of Information Studies, University of Copenhagen,
Copenhagen, Denmark

Abstract

The chapter takes an external view of the Italian evaluation experience in SSH, as described in various chapters of this volume. It compares the choices made by the Italian Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) to those of other countries, in light of the international literature on evaluation criteria, methodologies, and techniques. While not offering policy recommendations, the chapter articulates in a reflexive way the notion of research quality, and calls for a deepening of the cultural foundations of the evaluation exercise, based on cross-country differences rooted in language and history.

Keywords

Social science and humanities research; National research systems; National cultures; Scholarly research evaluation; Evaluation cultures

1. Introduction

Months before I started to write this chapter, I read an article titled: *11 Untranslatable Words from Other Cultures*. One of the words that appeared in this 'untranslatable' list was the Italian term: '*culaccino*'. This term may not be used often in Italy, but in an everyday context it simply means: "the mark left on a table by a cold glass" (Sanders 2013). *Culaccino*, amongst other internationally 'untranslatable' words (i.e., from Germany, Japan, Spain, Russia etc.) first appeared in a blog (Maptia.com), then later in Sanders (2013) news article, which was written to highlight a book: *Through the Language Looking Glass: Why the World Looks Different in Other Languages* (Deutscher 2010). The thesis of this book is captured by the following question: "If we hold language up as a mirror to the mind, what do we see reflected there: human nature or the cultural conventions of our society?" (p. 14).

According to Deutscher (2010), contemporary linguists tend to agree that "language is primarily an instinct" and "that the fundamentals of language are coded in our genes and are the same across the human race" (p. 6). Every international culture can impart its own labels onto things yet "the concepts behind these labels have been formed by the dictates of nature" (p. 13). In other words, humans generally observe things across nature in the same way, but speak or write about them differently. Because of this, "common sense would suggest that all cultures and all languages should be exactly the same" (p.7). Deutscher (2010) uses the example of words like "cat" and "dog",

which are so "clearly delineated by nature [that even though] children always need to be taught the labels for such concepts in a particular language of their society, they don't need to be told how to distinguish between the concepts themselves" (p. 12).

On the other hand, certain parts of a culture require further reflection, especially when a label is attributed to the realm of abstraction. Deutscher's (2010) second example is that "neither French nor German has a single concept, with a single label, that covers exactly the rank of meanings of the English concept of 'mind'. If you ask a bilingual dictionary to translate 'mind' into French, the dictionary will explain patiently that it depends on the context." (p. 14). A reverse situation would be the following: "English does not have a single concept that covers exactly the range of meanings of the French esprit. Again, a "dictionary would give a long list of different English words as possible translations" (p. 14).

In light of Deutscher's (2010) text, it can be useful to also think about scholarly research evaluation as a type of 'culture'. The evaluation culture, as we know it today, has evolved into a 'language' and this language includes many labels. Sometimes labels are given to fields of study (e.g., humanities, social sciences, natural sciences), or analytic methods (e.g., bibliometrics), and at other times they adhere to abstract concepts (e.g., quality; impact). Labels and concepts can also be influenced to a certain degree by the national context and language in which they are situated. In this sense the evaluation culture is a plural culture. In a country like Italy, this means that the way in which scholarly products are evaluated may not necessarily be transferable to all nations, but, the system itself can at least be observed at an international level (note: my discussions with a few authors from this volume occurred during the heat of the Italian summer; a natural time for everyone to observe culaccinos). The Italian culture is so greatly reflected by its language, that it will definitely influence how scholars from this country perceive and value academic research. This is especially true for research produced within the social sciences and humanities (SSH). Yet, to the Italian research community's credit they have provided us with an English translation of their work. My task; therefore, is to take this evaluative imprint; this metaphorical 'mark left by a cold glass on a table', and show how it relates to similar 'marks' left by other international scholars.

2. The Language of Scholarly Research Evaluation

In English, the language of scholarly research includes the following items for assessment: journals, articles, books, book chapters, reviews, etc. The Italian research community, of course, uses its own labels for these items, but each can be translated directly into English (as well as many other languages). In fact, most cultures that engage in scholarly research recognize similar, if not exactly the same objects related to specific labels. The technologies available to instantiate, preserve, and determine counts for these labeled items are also now practically universal (note: perhaps not always accessible, but at least universal). So, how can the Italian research system appeal to a general expectation of 'sameness', yet reveal something different? Three themes pertaining to research evaluation appear in this volume and each one has potential to unify or distinguish the Italian evaluation community from the international one. Notwithstanding Italy's economic and political structure, scholars from this country, like others across Europe, have become pre-occupied with the

following: (1) labeling and establishing classifications for scholarly outputs /academics /communities to be evaluated; (2) discovering or establishing useful datasets/peer communities to support the evaluation of these labeled constructs; and (3) defining and observing quality in scholarship and recognizing broad notions of impact.

I will examine specific chapters from this book under the microscope of the first two themes, and with the third theme, I will take the liberty of reflecting on some culturally informed ideas surrounding quality and impact.

2.1. Labels, Definitions, and Classifications

During a meeting with the editor of this volume, I was provided with a thorough introduction to the development of the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR). I will not describe this organization in detail, or compare it extensively to other national structures/organizations, unless it fits within one or more of my thematic discussions. Some of the individual chapters already provide an in-depth and comparative coverage of the Italian system vis-à-vis others. Here, the aim is to show how this volume is successful at examining and defining labels, in particular, those that have become a critical part of the Italian debate for the Social Sciences and Humanities (SSH).

2.1.1. Books

Sometimes when we identify *wh*at an item is not, we move closer to determining *wh*at it actually is. For example, in a previous text concerning SSH research evaluation, Hicks (2004) stated that the "danger of ignoring books is illustrated by exploring the differences between the worlds of book and journal publishing. Books are not just large journal articles" (p. 482). This is a logical assertion, yet Hicks (2004) does not elaborate further upon the publication industry or on the notion of what makes a book precisely more unique. Instead, she focuses on the metrics: "evidence is found in the lack of correlation between cites to books and journal articles" (p. 482).

Williams et al. (2018) have taken a more detailed approach to exploring what a book is, first by appealing to the reader's com*mon sense (e.g., How many pages are counted before a work is declared a book rather th*an an article?), then, by examining a definition of 'book' (e-book included), and noting that books are different from long reports and periodicals because they "do not qualify for an ISBN". Even with a useful definition, the authors show that there can still be anomalies and specifics. For instance, in some African countries, a book published by a scholar does not necessarily "adhere to the ISBN". Also, in Lithuania, where books do adhere to an ISBN, Williams et al. (2018) have found that for evaluation purposes, the item still has to be "4,000 characters multiplied by a field based coefficient 8".

When different languages are used to label and define a "book", is both the object and concept behind it basically the same? Yes and no. The basic label attributed to the object may be easy to translate, but the concept of *book* becomes more complicated when a decision has to be made about *wha*t to actually do with that book, and whether or not there is enough detailed record keeping (i.e., metadata for the book). The

evaluation culture dictates why, when, where, and how any given label is to be valued. Williams et al. (2018) use parts of their chapter to examine different national evaluation systems (e.g., the Dutch, British, Polish, Spanish, etc.) and conclude that qualitative and quantitative approaches each have unique drawbacks. Here, we see that the heart of the problem is not necessarily the approach, but how labels and definitions are established for a chosen approach. After "having summed up the situation in a number of European countries" the authors proclaim that "it is clear that no definition [of a book] can be applied everywhere".

Similar to other national systems, the Italians also have to rely on labels and definitions. In their chapter, Basili and Lanzillo (2018) define a book, or more specifically a monograph, in the following terms: a "broad, unified and articulated close examination" of a topic, which "cannot be decomposed into separate partitions". And, in distinct Italian style, they explain that "each element [of the monograph] helps to form the complex of the opera". Even if two authors are responsible for this composition, a monograph "cannot and should not be compared to a collection of chapters written by more than one person". At the ANVUR, this definition of book is broadened to include "book chapters and conference proceedings", all of which must have an ISBN. A book's scholarly purpose may also be characterized in terms of a "critical edition, translation, or scholarly commentary".

The unique part of Basili and Lanzillo's (2018) chapter; however, is not that they present us with yet another definition for a "book" but that they have also chosen to examine how the "same object can have a different role depending on the type of research it was meant to convey". This is highly pertinent, for example, to the field of legal studies, where many books fit the basic cri*teria* for what a book is (e.g., it possesses an ISBN), but cannot be taken at face value on the basis of genre. In law, certain books that are published as manuals might contain original work/research that is similar to what other research disciplines would expect from a traditional scholarly monograph (see Peruginelli and Faro, 2018).

2.1.2. Fields/Disciplines

Before the first evaluation took place in Rome at the ANVUR (i.e., the VTR – Valutazione Triennale della Ricerca) and prior to the development of other national evaluation systems, classifications were established for journals. Glänzel and Schubert (2003) note that journal classifications were originally created for retrieval purposes. Now, the scientific community relies on categories/classes for broad aspects of evaluation. In Denmark, for example, the classification system used to evaluate research does not follow the categories that we see for journals indexed in commercial databases (i.e., Clarivate Analytics' Web of Science). Instead, the Danish use an ad hoc approach, whereby scholars selected by the Danish Research Agency are invited to form peer groups (68 in total) representing an academic specialty. Each group works together to identify specialty-area journals that fit a level 1 (normal) or level 2 (prestigious) category. Level 1 journals are linked to a 1.0 reward point and level 2 are linked to 3.0 reward points (Giménez-Toledo et al. 2016). Approaches differ everywhere, and few classification schemes are alike, but many tend to be hierarchical, and each can be created using a *cognitive* approach, a *pragmatic* approach, or a *scientometric* approach when needed (see Glänzel and Schubert 2003).

The Italian evaluation system relies on a *cognitive* approach to classifying fields, and this includes 14 broad areas followed by sub-categories labeled with alphanumeric codes. For *example, area 11, designated as "Scienze storiche, filosofiche, pedagogiche e psicologich* (History, philosophy, pedagogy and psychology), includes "M-STO/01 STORIA MEDIEVAL (Medieval History) as well as "M-STO/04 M-STO/04 STORIA CONTEMPORANEA (Contemporary History). This cognitive classification has been useful for delineating between research outputs that can best be evaluated using bibliometrics (areas 1–9) versus those that benefit from assessments involving peer review (areas 10–14) (see Faggiolani and Solimine 2018). International scholars interested in the full Italian schema, known as the *Settori Scientifico-Disciplinari* (SSD) can find an English translation available online (see https://www.cun.it/uploads/storico/settori_scientifico_disciplinari_english.pdf).

Field labels are necessary for scholarly evaluation procedures, primarily for setting boundaries, and establishing areas for intensive study, but cultural conceptions behind them are open for debate. International scholars might examine, for instance, the degree to which History, Politics and Law constitute a natural cognitive grouping within the social sciences. Peruginelli and Faro (2018) suggest why it may not be a good idea: "law shares many of the peculiarities of social sciences since law is a social phenomenon, but when normativity and legal certainty are concerned, legal scholarship is probably closer to humanities". Also, an examination of the history of music might influence our perception of how well the field of musicology is suited to a quantitative-oriented evaluation program: Is music an art or is it a science? (Cohen 1984, 2010). The Arts and Humanities, or the Humanities and Social Sciences are frequently grouped, but some parts of social science are focused on human development and other areas are based on economic and political issues (Glänzel and Schubert 2003).

In this volume, the approach to field categorizations does not simply settle upon comparing the SSD to what has been introduced in other nations or at the European level (e.g., the ERC model). Faggiolani and Solimine (2018) emphasize why problems underlying evaluation systems are complex: "the correct classification of the branches of knowledge constitutes one of the key elements capable of fostering [all fields] with major cultural repercussions". While the authors are critical of the Italian SSD for being based primarily on an "administrative logic" as oppose to the "logic of scholarly communication" they show how this classification scheme can be transformed using a "specificity analysis". The objective of this analysis was to resolve conflicting points of administrative logic and communication logic, by aligning each product from a classified field, subfield, or level with a specific set of valuation criteria. Faggiolani and Solimine (2018) accomplished this using an automatic text analysis of terms used in a document written for the purpose of "identifying evaluation criteria and parameters for the SSH (areas 10-14). In this case, the text was the *Criteri, parametri e indicatori per l'Abilitazione scientifica nazionale*. With the results for the social science-like disciplines of history versus law, they have been able to identify key cultural-communication similarities and differences: (1) History (*monograph, continuity, law*) and (2) Law (*excellence, premise, monograph*).

2.1.3. Peer Review

The terms peer and review together propagate a host of challenges for the culture of scholarly research evaluation. European-based guidelines have been introduced to assist evaluators with peer selection procedures and setting up effective reviews. Capaccioni and Spina (2017) refer to these guidelines, but outline more specifically how the Italian community has initiated their own peer review procedure. n the 2011–2014 VQR (*Valutazione Della Qualita della Ricerca*) individual reviewers and committee reviewers were used to assess outputs from the SSH (Areas 10–14 in the SSD), with peer review defined in terms of "anonymous qualitative judgment". The review procedure focused on monographs, book chapters, research articles and other relevant outputs characterized by "innovation", "methodological rigor" and "proven or potential impact in the respective international scientific community" (see Capaccioni and Spina 2018).

Normally peer review is concerned exclusively with quality (i.e., often multi- ple notions of quality), but the alternative process of informed peer review gives individuals or committees the option of enlightening their judgments by observing quantitative data. When metric measures are introduced, peer review has potential to differ from its traditional counterpart. Scholars who are apprehensive about bibliometrics may fear that informed reviews will eventually disappear and that statistical indicators will take over. This fear is rooted in part by a mistrust in indicators in general, but is also reflects the built-in workloads, ethical issues and cost issues associated with recruiting and assigning scholars to reviewer roles. There are benefits, however, to keeping the evaluation culture open to informed peer review. For instance, peer experts/panel members comprised of both national and foreign scholars in Italy recently found that it can be problematic to expand the notion of quality to include internationalisation. Capaccioni and Spina (2018) argue that internationalisation is feasible as long as it "is intended as the ability of our disciplinary sectors to have a dialogue with the world of international research". But, they recognize also that not all scholarly outputs from the SSH will fit this ideal.

What we learn from the Italian experience is how critical it is to identify national products that fit within the realm of international research, "regardless of the lan- guage in which they are written" (see Capaccioni and Spina 2018). This is an open challenge to all national evaluation systems. In cases where internationalization is unclear, a procedure can be initiated to match each product from the SSH disci- plines with pre-determined levels for review. Conceptual levels help reviewers decide when quantitative data might serve as appropriate co-source of information for quality judgments. Capaccioni and Spina (2018) explain how the levels gradu- ate, starting with whether or not a SSH product is published in an international outlet, or translated to English or other languages, whether or not it is present in commercial databases (i.e., ISI-WoS/Scopus), if it has been co-authored by both a national and foreign author, produced as a result of funding by and international or European grant, and whether or not the product is included in alternative databases, like library catalogs.

3. Databases and Data Quality for the SSH

The international debate surrounding databases (or data types) suitable for evaluating SSH disciplines began a few years ago (Archambault and Vignola Gagné 2004; Gorraiz et al. 2013; Hicks and Wang 2009; Moed et al. 2009; Sivertsen and Larsen 2012); however, the Italian research community is not contributing too late to this discussion. This volume enters into a time period marked by a growing shift towards national data infrastructures for the SSH (see Giménez-Toledo et al. 2016), as well as an interest in Google Scholar, Google Books, and other under-exploited data- bases, like international library catalogs (Kousha et al. 2011; Torres-Salinas and Moed 2009; White et al. 2009). Here, I wish to highlight some of the techniques that the Italian researchers in this volume have used to retrieve, refine and utilize data for their most recent evaluation procedure, but first I want to start this section by examining what data means for the social sciences and humanities, and why it has to be approached with an open mind.

Data for the humanities scholar or the social scientist differs from data for disciplines across the sciences. Scientists collect data as they experiment with or observe natural phenomena. For the humanist, data is that which has been generated from the human mind (Bod 2013). Likewise, the social scientist works with data that is generated from human activity or human behavior. Statistical measures are applied to data quite often by social scientists, including scientific methods, but scholars from many humanities disciplines (e.g., philosophy; history; theology) prefer to write and reflect on data that is primarily textual and symbolic. This does not mean, however, that the products of humanistic research cannot be quantified. What it means is that when we work with textual and symbolic material *quantitatively* versus *qualitatively* there is potential to obtain different types and levels of insight (e.g., see Zuccala et al. 2014). In fact, the history of 'citation-ology' (i.e., the study of referencing and citation behavior) has already demonstrated that what we learn from approaching the highly *textual*, *contextual*, and *symbolic* citation, using a qualitative method of investigation can differ greatly than when we approach the same citation using a quantitative method of analysis (Bornmann and Daniel 2008; Brooks 1985; McCain 2006; Small 1978). When we grapple with measuring products and citations from humanistic research, it is less important; therefore, to distinguish data for specific forms of 'treatment' (e.g., *this data is qualitative/quantitative and that is not, so this should only be examined quantitatively/qualitatively*), and more important to focus on how the data needs to be curated to effectively support a chosen method.

In Great Britain, for instance, Thelwall and Delgado (2015) have already advo- cated for the humanities and humanities-based evaluations with a call out for "no metrics please, just data" (p. 817). The main argument put forth by these scholars is that data for evaluation procedures should not just be more plentiful, but enriched with contextual information. Consider now how a contextual effort might be made with respect to evaluating a scholarly monograph. A peer committee responsible for determining the monograph's quality (i.e., as an original, groundbreaking piece of scholarship) might choose to read excerpts from different chapters and/or search for some of the written reviews it has received. The inclusion of reviews then leads to the following question: was at least one review written about the book, or were none written at all? If the response to the first part is yes, the evaluation team has to decide if the monograph's quality should be linked symbolically to the binary *presence* of

public scholarly praise/criticism, or its *absence*. The general context then requires that the evaluators need to determine what this means at both a national or international level. But, the mere presence or absence of a review signifies only one thing, because when several reviews have been written, the number tallied together might produce yet another proxy of quality, and this in turn can bring to the assessment a point of comparison: should books with many reviews be valued or weighted more than those with few or no reviews? Now, even if there is a possibility to count reviews, the process does not have to stop there. Everything that is stated or written in the counted reviews can lead to yet another level of quality assessment because scholarly reviewers are invited to use their text to express, to some extent, both negative and positive opinions. Up to this point the evaluation procedure only gives attention to the content of the book and relevant forms of written recognition. Another proxy of quality could be the symbolic value (and weight?) that evaluators give to the number of citations received, or the publisher of the monograph. When the publisher serves as an indicator of quality this opens the debate up further to establishing notions of production quality, editorial quality, etc.

Since evaluation systems are not designed to focus on just one monograph, a procedure like this is, in reality, compounded when thousands of monographs are included. Context is important, but there is little room to incorporate all that is contextual when it is more critical to reduce the tension between what we expect data to look like and which databases currently record the most reliable data. In Italy, following the VQR 2004–2010, the recognition of and attempts at reducing this tension for the SSH have been linked to the following databases: Google Scholar and Library Catalogs.

3.1. Google Scholar

Google Scholar is receiving more and more attention as a tool for metric-based research evaluations but international researchers have been reluctant to use this resource without making comparisons to commercial indexes, mainly Scopus and Web of Science (e.g. Bar-Ilan 2008; Jacsó 2005; Meho 2007; Prins et al. 2016). The general consensus, following such comparisons, is that Google Scholar does not have a transparent data curation process, and that it tends to be "particularly poor in information" (Biolcati-Rinaldi et al. 2018). Metadata tags used for information such as type of document and language of contribution are absent from Google Scholar; hence Scopus and the Web of Science which both include more tags, support better opportunities for data retrieval (Biolcati-Rinaldi et al. 2018; Jacsó 2010). International scholars recognize, however, that one of the chief benefits of Google Scholar is its comprehensive coverage (Harzing and van der Wal 2008; Harzing 2014). With disciplines across the SSH, it can be particularly useful for examining books. For example, Kousha et al. (2011) extracted a sample of 1000 books submitted to the 2008 Research Assessment Exercise (RAE) in the United Kingdom, and found that Google Scholar produced "3.2 times more citations" than Scopus, including "medians [that were] three times as high" (p. 2147).

In Italy, the motivation for experimenting with Google Scholar stems from a Ministerial decree surrounding the methods that can be used at ANVUR to evaluate scientific outputs (from areas 1 to 9 in the SSD) versus outputs from the SSH (areas 10–14 in the SSD). Ferrara et al. (2018) note that a "distinction between bibliometric

and non-bibliometric scientific areas [was originally] defined by the Italian Ministry of University and Research (MIUR)". Thus far, the Italian evaluation system still recognizes the 'gold standard' of peer review, especially for the Social Sciences and Humanities, but this has not stopped a team of researchers from investigating whether or not data from Google Scholar might be reliable and valid for a real assessment procedure for certain subjects, such as Political Science, History, and General Sociology (i.e., subjects from the research area 14 of the SSD). Again, the work presented in this volume confirms that Google Scholar covers more data relevant to authors and outputs than Scopus or Web of Science (see Biolcati- Rinaldi et al. 2018). However, up to this point, this rate of coverage has never been assessed as thoroughly before without considering key differences between optimizing and cleaning data concerning authors from scientific research fields versus authors from the SSH. Because scientists co-author papers more frequently than scholars from the SSH, the potential of using co-author names in a disambiguation process is limited in these latter fields. This means that disambiguation techniques for author homonyms in the SSH need "solutions based on keyword and linguistic analysis" (see Ferrara et al. 2018).

In the follow-up study, which uses the disambiguated Google Scholar dataset, Biolcati-Rinaldi et al. (2018) present statistically significant correlation values resulting from a comparison of average peer review scores (for three scientific products submitted during the VQR 2004–2010) and two types of indicators extracted from Google Scholar, Scopus and Web of Science (i.e., h-index values and citations per contribution values). Here we have some insight as to the degree to which a relationship exists between the *peer recognized quality* of a research product and its *indicators of scientific recognition*. The Italian system could be leaning toward operationalizing informed peer review for certain disciplines in the future, but Biocalti-Rinaldi et al., indicate that it is still important to "keep distinct the bibliometric level from the evaluative one" because they see the "two disciplines as relatively autonomous".

3.2. Library Catalogs

In 2009, two teams of researchers from different parts of the globe simultaneously suggested that library catalogs might be a good source of data for evaluating book-oriented fields across the SSH (Torres-Salinas and Moed 2009; White et al. 2009). Prior to 2011, which was the year when Thomson Reuters (now Clarivate Analytics) Book Citation IndexTM was finally introduced, data resources for books were especially poor and bibliometricians simply could not rely on journal citation indexes alone for evaluating SSH outputs. Torres-Salinas and Moed (2009) decided, therefore, that one could try to "use the same tools and methods as those applied by bibliometricians" but focus on data from library catalogs. Their underlying rationale was that the "inclusion of a book in academic libraries [could be seen] as an expression of its utility for the academic community" (p. 2). An analogy could be drawn "between traditional citation analysis of journal articles and a Library Catalog Analysis of book titles" (p. 2). At the same time, White et al. (2009) advanced this analogy by coining the term "*libcitation*". According to this research group, the libcitation possesses a symbolic value that is quite similar to the citation: "when librarians commit scarce resources to acquiring and cataloging a book, they are in

their own fashion citing it, just as scholars do when they refer to it in new works of their own; both are engaged in bibliographic speech acts" (p. 1084).

To date, a handful of studies have grown from these two contributions, and for researchers with a growing interest in library holding counts, the OCLC-WorldCat seems to be the international catalog of choice (Linmans 2010; Kousha et al. 2017; Zuccala and White 2015). OCLC-WorldCat is a union catalog, and because it currently covers libraries worldwide – i.e., academic libraries, national libraries and public libraries – much can be said about a book's perceived *cultural visibility* and *utility*. The libcitation symbolizes the cultural and perceptual element; while the citation can corroborate perception by serving as a symbol of utility. For this reason, it can be interesting to compare the number of academic/institutional libraries that hold a book worldwide, to the number of citations it has received (or did not receive after it was published and cataloged (see Zuccala and White 2015). According to Zuccala and White (2015), "both citations and libcitations exhibit the highly skewed distributions typical of bibliometrics" (p.308). Yet, in a test of the relationship between citations and libcitations, Spearman correlation results show that none "are strong enough to indicate that libcitations can substitute for citations as a measure" (p.309).

Biagetti et al. (2018) have taken an altogether different approach to studying library catalogs, and their focus is on the question of data reliability. Here, a suggestion is made that the libcitation may not be the most reliable indicator of a book's esteem [as per White et al. 2009's argument], primarily because the underlying process – i.e., the book's acquisition process – cannot always be verified. Biagetti et al. (2018) remind us that many libraries have pre-approved acquisition plans set up with top publishers. Also, many books are purchased by librarians with special attention and intention; while others appear in the same catalog because they have been donated, or sent to the library as a gift. One could survey a sample of libraries to ascertain these underlying factors, as Biagetti et al. (2018) suggest. Or, one could simply accept White et al.'s (2009) approach, which is to trust the role of librarians: "libcitations reflect librarians' knowledge of audiences...what librarians know about the prestige of publishers, the opinions of reviewers, and the reputations of authors" (p. 1084). There is no 'correct' answer to this problem; thus, a refusal to take the libcitation at face value is just similar to questioning the validity of counting citations – i.e., it is not particularly easy to ascertain the reason for a citation either (note: think citer motivation studies and citation context studies).

4. Some Reflections on Scholarly Quality and Impact

4.1. Culture and Perceptions of Quality

In the introductory section of this chapter I referred to a book written by Guy Deutscher (2010), titled *Through the Language Looking Glass: Why the World Looks Different in Other Languages* and suggested that scholarly research evaluation is a type of culture (a pluralistic culture, in fact). I would like to come back to this point, and focus on some of the definitions that we can give to abstract terms like *quality* and *impact*. Within Europe it might seem ideal if every country could agree upon definitions for these interrelated terms, especially with respect to scholarly research,

but this may be quite problematic. Not only would this defy Deutscher (2010)'s thesis concerning *why the world looks different in other languages*, it also negates the fact that much of what is produced by international scholars from the social sciences and humanities is *supposed* to be cultural, or at least situated within a cultural context.

The word *culture* is frequently used in the field of anthropology, and for most anthropologists it is defined in terms of "the shared set of (implicit and explicit) values, ideas, concepts, and rules of behaviour that allow a social group to function." It is also understood to be "dynamic and evolving" because it is the "constructed reality that exists in the minds of social group members" (Hudelson 2004). The Italian culture, for example, has many different phrases or terms that can be used to evoke a shared idea of *quality*. One such term, which is not easy to translate directly into English, is *sprezzatura*. Baldassare Castiglione, the Italian courtier, diplomat, soldier and Renaissance author, was the first person to coin this term because he believed that it captured best what it meant to be a courtier. The ideal Rennaisance courtier was expected to be skilled at many things (i.e., athletics, music, dancing etc.) but somehow he was also expected to be quite modest. Rebhorn (1978) defines this performance-oriented quality as the ability to display "an easy facility in accomplishing difficult actions", yet "hide the conscious effort that went into them" (p. 33).

Another modern phrase for recognizing quality in Italy is *regola d'arte*. In English it means that something has been produced "in accordance with the state of the art". Normally, an Italian is more likely to use this word to show appreciation for an object crafted by a workman (e.g., a lamp), not a piece produced by a scholar (e.g., a book), but we can still see from this country's language that in the minds of the Italians, something of high quality is that which has been approached "artfully"; thus quality is analogous to being artful.

Since culture is expressed through language, the Italian terms *sprezzatura* and *regola d'arte* indicate how *quality* is a term that has potential to be quite contextually and culturally specific. These examples also demonstrate the degree to which they are time dependent. During the Renaissance period the word *sprezzatura* had a much more positive connotation than it does today. Now it would be taken to mean that a person possesses "an ostentatiously nonchalant attitude of studied indifference" (English translation derived from http://www.treccani.it/vocabolario/sprezzatura/). Modern Italians are thus more likely to recognize a quality performance using other terms. How then can we expect to create a unified notion of scholarly quality, when the international culture of research evaluation is not really a singular culture? Scholarly quality has to reflect a cultural orientation, and naturally different notions of *quality* will evolve over time.

4.2. The Problem of 'International' Impact

In the *Evaluation Society*, Peter Dahler Larsen's (2012) approach to evaluation was to "illustrate what [it] looks like when it is culturally compatible with reflexive modernization. For Larsen (2012), evaluation has become "the manifest sign" that an organization [and even a nation] "is capable of adapting itself to changing conditions" and that it is falling "in line with present cultural expectations" (p. 144).

While the Italian research community has its own language for perceiving and recognizing *quality*, scholars today are under pressure to adapt to changing conditions produced by the international evaluation culture; conditions that require them to produce texts that have (or can potentially have) an international impact. One of the key questions under debate is whether or not a scholarly product has to be produced in or translated to English in order for this to occur (see Capaccioni and Spina 2018). Thus, similar to the concept of *quality*, the word *impact* needs to be examined quite carefully. If the language of a given culture can play a significant role in the recognition of *scholarly quality*, translation theory can help us understand the problem of international impact.

According to Pellizzi (2012), there are two main approaches to translation: one that is *literary* and another that is *scholarly*. A *literary* translation is focused on the target language culture and aims to produce the best analog of the source language text at the time of translation. For a scholarly text, this would entail that the translator belongs to the source language and target language culture, as well as the research domain for which the translation is needed. In this sense, both language and culture not only refer to a country or region and its societal and linguistic distinctiveness, but the scholarly community or tribe within that country/region, which also possesses values, traditions, terminologies, etc. in its own right. The pitfall related to this form of translation is that the translator might to a greater or lesser degree subjectively re-interpret or alter the original meaning of a text (either knowingly or unknowingly) according to his or her own view of the research domain. Pellizzi (2012) points out that especially for non-fiction texts literary translators run the risk of "going native", or not actually producing an exact translation of the original, but instead making an adapted free translation of the original text.

The *scholarly* translation is focused on the source language text and the cultural-intellectual and historical context in which it was produced (Pellizzi 2012). Because it is different from the *literary* approach to translation, being fluent in a language and having an insider's understanding of a culture is of lesser importance and might even be an advantage. A translator with an outside view has potential to remain true to the original text and thus create a translation that is considered more objective. However, this does not mean that this approach is complete free from complications or consequences. The implied objectiveness of a scholarly translation is that the translator has to make certain that the reader is always aware of the gap between the intellectual world of the original author and the new translation. If a translator is disconnected from a full understanding of both the cultural and linguistic subtleties behind the text, meaning may get lost or distorted somehow in the translation process.

The message that we get from from Pellizzi (2012) is that when we translate an original language text into another language there is the potential risk of altering its original inherent *quality*. Bruno Bettleheim's (1983) book, titled *Freud and Man's Soul* provides the perfect example:

> *When in middle age, I was fortunate enough to be permitted to start a new life in the United States, and began to read and discuss psychoanalytic writings in English. I discovered that reading Freud in English translation leads to quite different impressions that I had formed when I read him in German. It became*

> *apparent to me that the English renditions of Freud's writings distort much of the essential humanism that permeates the originals" (p. 3).*

Bettleheim continues to explain that

> *[in] his work and in his writings, Freud often spoke of the soul- of its nature and structure, its development, its attributes, how it reveals itself in all we do and dream. Unfortunately nobody who reads him in English could guess this because nearly all his many references to the soul, and to matters pertaining to the soul have been excised in translation... [The] most important and original concepts of psychoanalysis, makes Freud's direct and always deeply personal appeals to common humanity appear to readers of English as abstract, depersonalized highly theoretical, erudite, and mechanized—in short, scientific statements about the strange and very complex workings of our mind (p. 5).*

Since the humanistic efforts of Freud have at one time been altered to suit the intellectual palate of American scholars, and since history has a tendency to repeat itself, evaluators might try to establish qualitative and quantitative assessment procedures that are much more responsive to the problem of 'internationalised' impact. Social scientists and humanities scholars should not be pressured towards producing outputs that have 'potential' for international impact, if it does not reflect their cultural sensibilities. Instead, the system should be designed to guide them with critical information about when it is worthwhile to focus on translating their own works, or have them translated by others for international readers. The translation process itself does not necessarily have to be problematic, as long as it takes a 'clean' *literary* or a 'clean' *scholarly* approach (note: a new approach to clean data!). Moreover, texts that hold a certain meaning and quality within a particular cultural context might perhaps be evaluated separately from their translated versions, as well as documents intended to be of international interest at the outset. This is one (suggested) way to establish a pluralistic culture of evaluation; it may in fact be the only way to respect the value of humanistic research.

5. Conclusion and Acknowledgments

After the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) was first established in 2006, many Italian scholars have made painstaking efforts to improve upon various methods of evaluation related to scholarly communication practices and outputs across in Italy. Since I have taken a fairly broad approach to traversing the current evaluation landscape, it is important to note that I did not cover every issue that needs to be addressed, or every issue that was highlighted within this particular volume. I would therefore like to urge readers to give special attention to each individual chapter, and treat the present chapter as a useful summary, based on a specific thematic intention.

I would also like to mention that I am not a policy-maker and do not wish to make any official recommendations to that effect. I do believe; however that future evaluation practices should not be hindered due to a lack of resources, fear, ethical issues (which can be monitored), or cultural comparisons that might leave one country

believing that its approach to evaluation is better than another. The plural evaluation culture that I describe in this chapter – a European-Anglo culture at this point – possesses greater opportunities for uniting countries/nationalities over a respect for differences, rather than forced ideals. Each country has its own 'mark' to leave and each can have some influence, but no mark should be so permanent that we are locked into a system that cannot evolve.

I wish to thank Andrea Bonaccorsi for giving me the opportunity to learn a great deal about the Italian research evaluation system and for trusting me with the preparation of this chapter. My gratitude also goes to Ginevra Peruginelli for hosting me at her home and at the ITTIG in Florence, at a time when it was crucial to attend a meeting concerning this volume. Alfio Ferrara, Ferruccio Biolcati-Rinaldi, and Luca Lanzillo also attended this meeting, and I am grateful to each of them for sharing their research experiences with me and for engaging me in a lively discussion.

References

Archambault, E., & Vignola Gagné, E. (2004). The use of bibliometrics in the social sciences and humanities. Science-Metrix, Final Report. Prepared for the Social Sciences and Humanities Research Council of Canada (SSHRCC).

Bar-Illan, J. (2008). Which h-index? – A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74(2),* 257–271.

Basili, C., & Lanzillo, L. (2018). Research quality criteria in the evaluation of books. In this volume.

Bettleheim, B. (1983). *Freud and man's soul. An important re-interpretation of Freudian theory*. New York: Vintage Books.

Biagetti, M. T., Iacono, A., & Trombone, A. (2018). Is the diffusion of books in library holdings a reliable indicator in research assessment? In this volume.

Biolcati-Rinaldi, F., Molteni, F., & Salini, S. (2018). Assessing the reliability and validity of Google Scholar indicators. In this volume.

Bod, R. (2013). *A new history of the humanities: the search for principles and patterns from antiquity to the present*. Oxford: Oxford University Press.

Bornman, L. & Daniel, H.-D. (2008). What do citations measure? A review of studies on citing behaviour. *Journal of Documentation, 64(1),* 45-80.

Brooks, T. A. (1985). Private acts and public objects. An investigation of citer motivations. *Journal of the American Society for Information Science and Technology, 36(4),* 223-229.

Cohen, H. F. (1984). *Quantifying music. The science of music at the first stage of the scientific revolution. 1580 to 1650*. Dordrecht, NL: Reidel.

Cohen, H. F. (2010). Music as science and as art. In R. Bod, J. Maat, & T. Westeijn (Eds.) *The Making of the Humanities. Volume I. Early Modern Europe*. Amsterdam NL: Amsterdam University Press.

Deutcher, G. (2010). *Through the language looking glass: Why the world looks different in other languages*. New York: Metropolitan Books.

Faggiolani, C., & Solimine, G. (2018). Mapping the role of the book in evaluation at the individual and department level in Italian SSH: A multisource analysis. In this volume.

Ferrara, A., Montanelli, S., & Verzillo, S. (2018). Google Scholar as a citation database for non- bibliometric areas: The EVA project results. In this volume.

Giménez-Toledo, E., Manana-Rodrıguez, J., Engels, T. C. E., Ingwersen, P., Polonen, J., Sivertsen, G., Verleysen, F. T., & Zuccala, A. A. (2016). Taking scholarly books into account. Current developments in five European countries. *Scientometrics, 107(2)*, 685–699.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56(3),* 357–367.

Gorraiz, J., Purnell, P., & Glänzel, W. (2013). Opportunities and limitations of the Book Citation Index. *Journal of the American Society for Information Science and Technology, 64(7),* 1388–1398.

Harzing, A.-W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics, 98(1)*, 565–575.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed et al. (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473–496). The Netherlands: Kluwer Academic Publishers.

Hicks, D., & Wang, J. (2009). Towards a bibliometric database for the social sciences and humanities. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1017&context=diana_hicks http://eproofing.springer.com/books/printpage.php?token=zICJjtpgBqkIjPLfbnL-RGDLD3UwwWV4ZFhzvFc6kbM 11/13.

Hudelson, P. M. (2004). Culture and quality: An anthropological perspective. *International Journal for Quality in Health Care, 16(5),* 345–346.

Jacsó, P. (2005). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science, 89(9),* 1537–1547.

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review, 34(1)*, 175–191.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology, 62(11),* 2147–2164.

Kousha, K., Thelwall, M. & Abdoli, S. (2017). Goodreads to assess the wider impact of books. *Journal of the Association for Information Science and Technology, 68(8),* 2004-2016.

Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics, 83(2),* 337–354.

McCain, K. W. (2006). How influential is Brooks' Law? A longitudinal citation context analysis of Frederick Brooks' The Mythical Man-Month. *Journal of Information Science, 32(3)*, 277–295.

Meho, L. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58(13),* 2105–2125.

Moed, H.F., Linmans, J., Nederhof, A, Zuccala, A., Lopez Illescas, C., & de Moya Anegon, F. (2009). Options for a comprehensive database of research outputs in the Social Sciences and Humanities. Research report to the Project Board of the Scoping Study "Towards a Bibliometric Database for the Social Sciences and Humanities". Standing Committees for the Social Sciences and the Humanities of the European Science Foundation (ESF).

Peruginelli, G., & Faro, S. (2018). Research quality evaluation: The case of legal studies. In this volume.

Prins, A. A. M., Costas, R., van Leeuwen, T., & Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation, 25(3),* 264-270.

Rebhorn, W. A. (1978). *Courtly performances: Masking and festivity in Castiglione's book of the courtier.* Detroit: Wayne State University Press.

Sanders, E. F. (2013, October 26). 11 untranslatable words from other cultures. Huffington Post. Retrieved from http://www.huffingtonpost.com/ella-frances-sanders/11-untranslatable-words-f_b_3817711.html.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91(2),* 567–575.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*, 327–340.

Thelwall M. & Delgado, M. M. (2015). Arts and humanities research evaluation: No metrics please, just data. *Journal of Documentation, 71(4*), 817–833.

Torres-Salinas D. & Moed, H. F. (2009). Library catalog analysis as a tool in the studies of social sciences and humanities. An exploratory study of published book titles in economics. *Journal of Informetrics, 3(1),* 9–26.
White, H., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for the comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60(6),* 1083–1096.

Williams, G., Basso, A., Galleron, I., & Lippiello, T. (2018). More, less or better: The problem of evaluating books in SSH research. In this volume.

Zuccala, A., & White, H. D. (2015). Correlating libcitations and citations in the humanities with WorldCat.org and Scopus Data. In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of the 15th International Society for Scientometrics and Informetrics (ISSI),* Istanbul, Turkey, 29th June to 4th July, 2015. (pp. 305–316). Bogazici University.

Zuccala, A., van Someren, M., & van Bellen, M. (2014). A machine-learning approach to coding book reviews as quality indicators: Towards a theory of mega-citation. J*ournal of the American Society for Information Science and Technology, 65(11),* 2248-2260.