
Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias

A preliminary comparative study across disciplines in English, Spanish and Arabic

- > Imogen Casebourne
- > Dr. Chris Davies
- > Dr. Michelle Fernandes
- > Dr. Naomi Norman

Epic



UNIVERSITY OF
OXFORD



Casebourne, I., Davies, C., Fernandes, M., Norman, N. (2012) Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias: A comparative preliminary study across disciplines in English, Spanish and Arabic. Epic, Brighton, UK. Retrieved from: http://commons.wikimedia.org/wiki/File:EPIC_Oxford_report.pdf

This work is licensed under the *Creative Commons Attribution-ShareAlike 3.0 Unported Licence* <http://creativecommons.org/licences/by-sa/3.0/>

Supplementary information

Additional information, datasets and supplementary materials from this study can be found at: http://meta.wikimedia.org/wiki/Research:Accuracy_and_quality_of_Wikipedia_entries/Data

Table of Contents

Table of Contents.....	2
Executive Summary.....	5
1. Introduction	5
2. Aims and Objectives.....	5
3. Research Methodology	5
4. Data Coding and Analysis.....	6
5. Results.....	6
6. Discussion.....	7
Acknowledgements.....	8
1. Introduction	9
2. Aims and Objectives.....	13
3. Research Methodology	14
3.1 Selection Criteria.....	14
3.2 Sampling.....	18
3.3 Selection of articles.....	19
3.4 The Review Process.....	20
4. Data Coding and Analysis.....	26
4.1 Data Coding.....	26
4.2 Quantitative Analysis	27
4.3 Qualitative Analysis.....	28
5. Results.....	29
5.1 Quantitative Analysis	29
5.2 Qualitative Findings	38
6. Discussion.....	50
6.1 Methodology.....	50
6.2 Findings	53
6.3 Recommendations	55
Appendix I	59
1. Prototype for Selection of Spanish Encyclopaedias for Comparison with Spanish Wikipedia	59
2. Student Biography and Nomination Sheet	59
3. Article Feedback Questionnaire.....	59
4. Article Comparison Questionnaire.....	59

5. Declaration..... 59

6. Example of Article Ready For review, After Completion of Standardisation and Anonymisation
Process 59

Executive Summary

1. Introduction

Previous studies, most notably the one carried out by the journal *Nature* in 2005, have sought to compare the quality of *Wikipedia* articles with that of similar articles in other online Encyclopaedias. In part as a result of the findings of such studies, *Wikipedia* has instigated a number of processes for assessing the quality of its entries, inviting readers and editors to rate articles according to criteria such as trustworthiness, neutrality, completeness and readability. Recently, *Wikipedia's* founder Jimmy Wales highlighted the value of conducting a study which analysed articles across both languages and subjects to allow differences in levels of accuracy and quality across language and subject domains to be identified. The results could inform editor recruitment efforts and the design of expert feedback mechanisms.

The size, scope and complexity of undertaking such a large-scale study necessitated gathering preliminary evidence to inform the methodology and design. It was therefore decided that a small-scale preliminary project would be essential to determine a sound research methodology, which is the reason that the present pilot study was undertaken. The present study, funded by the Wikimedia Foundation, presents the background, methodology, results and findings of a preliminary pilot conducted by Epic, a UK-based e-learning company, in partnership with the University of Oxford.

2. Aims and Objectives

The key aims of this pilot study are as follows:

1. To explore the opinion of expert reviewers regarding attributes relating to the accuracy, quality and style of a sample of *Wikipedia* across a range of languages and disciplines.
2. To compare the accuracy, quality, style, references and judgment of *Wikipedia* entries as rated by experts to analogous entries from popular online alternative encyclopaedias in the same language.
3. To explore the viability of the methods used in respect of the first two aims for a possible future study on a larger scale.

3. Research Methodology

Three languages were selected for study: English, Spanish and Arabic. Pairs of articles in those languages were selected in the following broad disciplinary areas: (a) Humanities, (b) Social Sciences, (c) Mathematics, Physics and Life Sciences and (d) Medical Sciences. Each pair consisted of an article from *Wikipedia*, and an article from one of a range of comparator online encyclopaedias: *Encyclopaedia Britannica* (English), *Enciclonet* (Spanish), *Mawsoah* and *Arab Encyclopaedia* (Arabic).

Twenty-four postgraduate students of the University of Oxford were selected to help review pairs of articles and to identify academic experts in their fields who would be recruited to review the same pairs of articles. Thirty-three academic experts were finally recruited. All

possessed doctorates and were employed in academic posts at a highly rated department within a well-established university. All students and academic experts were fluent in the target languages.

A feedback tool was devised for eliciting numerical scores and qualitative comments about the articles, which were reviewed blind by the academics, who were asked to certify that they had not sought out the original articles online during the review process. The feedback tool provided academics with a wide range of quality criteria, drawn from extensive previously published research.

Articles were standardised so as to erase information which helped to identify their origins; in particular, checks were carried out to ensure that a particular article was not the victim of vandalism (although this did not impact on article selection for the present study).

Twenty-two articles were selected in all. Some difficulty was encountered in finding articles of sufficient substance and scope in encyclopaedias paired with *Wikipedia* in different languages.

4. Data Coding and Analysis

Quantitative and qualitative data were analysed through separate processes. Quantitative data analysis was carried out on the sample overall, in relation to each language separately, and in relation to each disciplinary area separately. Data was coded in five main dimensions: i) accuracy, ii) references, iii) style/ readability, iv) overall judgment (including citability), v) overall quality score.

Qualitative analysis was initially carried out blind, and involved the reduction and display of reviewers' comments so that these could be compared with one another, in relation to specific articles, pairs of articles and across the sample as a whole. The qualitative analysis aimed to capture both the opinions of reviewers about specific aspects of the articles, and their overall judgments about each individually and in comparison with the other in the pair.

5. Results

All of the results outlined below are based on a small sample studied for the purposes of piloting the study's approach and methods, and these results cannot therefore be generalised to the wider output of the online encyclopaedias referred to.

Quantitative results for the articles reviewed show that the *Wikipedia* articles in this sample scored higher overall than the comparison articles with respect to accuracy, references, style/ readability and overall judgment. The scores for the latter item, which includes citability, indicated that none of the encyclopaedias were rated highly by academics in terms of suitability for citation in academic publications.

Results across languages showed that *Wikipedia* fared well in this sample against *Encyclopaedia Britannica* in terms of accuracy, references and overall judgement, but no better on style and overall quality score. The same was true of *Enciclonet*, but the Arabic encyclopaedias scored significantly higher on style than *Wikipedia* and equally well on the other criteria.

Results across disciplines showed that *Wikipedia* scored higher in this sample in terms of provision of references in humanities-based articles, but no differences were apparent in terms of the other criteria, as was also the case with articles in mathematics, physics and life sciences. There was a similar result for articles in social sciences, but with higher scores on style/ readability for the other encyclopaedias. In medical science articles, *Wikipedia* scored significantly higher on accuracy, references and overall judgment, but there were no differences on the other criteria.

Qualitative results for this sample showed similar findings, but also revealed the importance to reviewers of articles possessing a sense of cohesiveness and structure. Although many *Wikipedia* articles in the sample were commented on favourably, they were criticised in some cases for lacking cohesiveness and for internal inconsistencies and repetition. Reviewers were particularly approving of articles that presented an engaging and coherent introduction to a topic, rather than excessive amounts of information.

The same differences seen in the quantitative analysis were evident in the qualitative with respect to different languages. In terms of different disciplines, small differences in terms of favoured quality criteria were evident, such as an emphasis on the notion of conciseness in the science-based article reviews.

6. Discussion

In many respects, the methodological approach had proved productive and workable on the small scale of the present study. But it was recognised that there were difficulties (even on this small scale) in terms of identifying appropriate articles, recruiting a sufficient range of reviewers, and anonymising articles which, if the study were to be carried out on a far larger scale, would possibly prove hard to surmount. Therefore, it is recommended that the viability of a larger study of this kind in the future should be considered cautiously, and that consideration might be given instead to carrying out a series of more compact studies of this kind over time.

It is also recommended that more research might be carried out on what is reasonable and appropriate to expect of online encyclopaedia content. It was clear from this study that, while many academics spoke in positive terms about a high proportion of articles reviewed from all encyclopaedias, it was not the case that they were inclined to regard these as being citable in academic publications alongside peer-reviewed journals and published books. We recommend that more research is done on how users interpret and make sense of content from online encyclopaedias in general and from *Wikipedia* in particular.

Overall, the *Wikipedia* articles in this very small sample, investigated as part of a pilot study only in this instance, fared well in comparison with articles from other encyclopaedias. While no generalisations can be made from this outcome, these findings do help to point researchers in future studies towards investigation of the unique qualities of *Wikipedia*, as a source of knowledge that was shown in the small number of instances studied here at least to be capable of producing articles that were markedly up to date and well referenced.

Acknowledgements

The investigators of this study are grateful to all the academic and student experts from all parts of the world who participated in the study and undertook the reviewing of the articles with great rigour. The investigators gratefully acknowledge those who helped in the selection of the encyclopaedias and articles in Arabic and Spanish, and in the standardisation and anonymisation of the articles before review.

The investigators are thankful to the Clarendon Scholars Council, the Spanish and Latin American Societies, the Middle East Centre and the Khalil Research Institute, the Medical Sciences Division and the Middle Common Rooms of various colleges of the University of Oxford for their help in recruitment.

The investigators appreciate all those students and academics at the University of Oxford who expressed an interest in the study and thank them all for their overwhelming enthusiasm.

The investigators are grateful to the Department of Education at the University of Oxford and to Exeter College for hosting various aspects of the study.

The investigators are grateful to the Wikimedia Foundation for granting the funding to carry out this study.

1. Introduction

The popularity of online encyclopaedias as a source of information has increased tremendously in the past two decades. However, the issue of the quality and accuracy of the information available in online encyclopaedias remains one of debate. This is particularly the case in those encyclopaedias available on the internet which do not charge users to access information. There has, however, been much discussion about the accuracy of information available in 'free' online encyclopaedias, which do not pay contributors and editors a fee but instead rely on voluntary contributions from persons who regard themselves experts without formal clarification of their qualifications or a stringent process of peer-review or editing. While this characteristic facilitates rapid and free transfer of knowledge, critics argue that 'opening the editing process to all regardless of expertise means that reliability can never be ensured'¹.

According to the leading global provider of web metrics, Alexa.com, *Wikipedia* is the most popular online encyclopaedia and the sixth most popular website in the world¹. It has more than 19 million articles in 270 languages. All content is freely available and approximately 13-15% of global internet users visit *Wikipedia* each day. *Wikipedia* is a collaboratively compiled and edited encyclopaedia with contributions in the form of text, pictures, formatting, citations and lists from multiple, unpaid editors and professionals. The process is regulated by means of an explanation of changes made between editors, notability guidelines and a tutorial process for new editors. Disputes about content are usually resolved by discussions between '*Wikipedians*', i.e. users, contributors and editors.

In December 2005 the scientific journal *Nature* reported on a study they had undertaken to compare the accuracy of science entries on *Wikipedia* with those on the online version of *Encyclopaedia Britannica*². Unlike *Wikipedia*, which relies on voluntary contributors, regardless of proven mastery or qualifications, *Encyclopaedia Britannica* uses selected paid expert advisors and editors. At the time of the *Nature* study, *Wikipedia* comprised 3.7 million articles in 200 languages and was ranked the 37th most visited website on the internet².

Nature invited independent academic scientists to peer review entries (in the English language) for their particular areas of science expertise, from both *Wikipedia* and *Encyclopaedia Britannica*. Each scientist was asked to identify any inaccuracies and comment on the articles' quality and readability, without being aware of the source of the article. Forty-two reviews were submitted to *Nature* revealing on average four inaccuracies per *Wikipedia* article, in contrast to three per *Encyclopaedia Britannica* article. The general response was one of surprise, with levels of accuracy in *Wikipedia* being better than expected. *Wikipedia* articles were rated more 'poorly structured and confusing' compared to articles from *Encyclopaedia Britannica*, with 'undue prominence being given to controversial scientific theories'². Nevertheless, for *Encyclopaedia Britannica*, the oldest continuously published reference work in the English language, the results were worse than

¹ <http://www.alexa.com> (April 2012) *Top Sites*, [Online], Available at: <http://www.alexa.com/topsites> [Accessed 12/04/12].

² Giles, J. (2005) 'Internet encyclopaedias go head to head', *Nature*, vol.438, 15 December 2005, pp. 900-901.

expected². While Jimmy Wales, the co-founder and promoter of *Wikipedia*, expressed delight, he also added: “Our goal is to get to *Britannica* quality or better”¹.

In a rebuttal published in 2006, *Encyclopaedia Britannica* refuted *Nature’s* findings, stating: ‘Almost everything about the journal’s investigation, from the criteria for identifying inaccuracies to the discrepancy between the article text and its headline, was wrong and misleading’³. The rebuttal stated that the conclusion of *Nature’s* report was false, because the journal’s research was invalid and clearly stated that the purpose of its production was to ‘reassure *Britannica’s* readers about the quality of our (*Britannica’s*) content, and to urge that *Nature* issue a full and public retraction of the article’³. The document highlighted a number of concerns about *Nature’s* research methodology³ including:

1. The lack of availability of the reviewers’ reports.
2. The selection of *Britannica* articles in an unstandardised manner from productions of the encyclopaedia (such as *Britannica Student Encyclopaedia* and *Britannica Book of the Year*) rather than solely from *Encyclopaedia Britannica*.
3. The selection of only parts and sections of *Britannica* articles rather than entire entries.
4. Rearrangement and re-editing of *Britannica* articles for the purpose of the study, including the merging of passages from two separate articles.
5. Failure to clarify the factual assertions of the reviewers.
6. Lack of distinction between minor inaccuracies and major errors.
7. Clarification that the reviewers’ comments were based on facts and not opinions.
8. Misinterpretation and misleading presentation of the results.

Nature responded by rejecting *Encyclopaedia Britannica’s* criticisms, affirming its confidence in the study, and refusing to retract⁴. Numerous other non-academic and academic publications have followed *Nature’s* example, yielding interesting results. In 2007, a study by *Stern* magazine⁵, compared 50 articles from the *German Wikipedia* to *Brockhaus Enzyklopädie*⁶, the largest German language printed Encyclopaedia in the 21st century. Fifty articles from disciplines spanning politics, business, sports, entertainment, geography, science, medicine, history, culture and religion were rated by experts for accuracy, completeness, timeliness and clarity. *Wikipedia* achieved a mean overall score of 1.7 across disciplines on a scale from 1 (best) to 6 (worst), while entries for the same keywords from the paid online edition of the 15-volume *Brockhaus* achieved an average overall score of 2.7. *Wikipedia* articles scored higher on timeliness and accuracy than articles from *Brockhaus Enzyklopädie*, although the *Wikipedia* articles were judged too complicated for a lay audience.

The accuracy of *Wikipedia* entries in the sciences has been scrutinised. In a study published in the *Annals of Pharmacotherapy* in 2008, Clauson and colleagues found the scope, completeness and accuracy of drug information in *Wikipedia* to be statistically lower than

³ Encyclopædia Britannica, Inc. (March 2006), *Fatally flawed: refuting the recent study on encyclopaedic accuracy by the journal Nature*, [Online], Available at: http://corporate.britannica.com/britannica_nature_response.pdf [Accessed 11/03/11].

⁴ Nature (23 March 2006), *Encyclopaedia Britannica and Nature: a response*, [Online], Available at http://www.nature.com/press_releases/Britannica_response.pdf [Accessed 11/03/11].

⁵ <http://www.stern.de/digital/online/stern-test-wikipedia-schlaegt-brockhaus-604423.html>

⁶ <http://www.brockhaus.de/enzyklopaedie/30baende/index.php>

that in a free, online, traditionally edited database (Medscape Drug Reference [MDR])⁷. In a report establishing the internal validity of *Wikipedia* entries for 39 of the most commonly performed inpatient surgical procedures in the U.S., 100% presented accurate content while 85% of the entries contained appropriate information for patients⁸. Interestingly, there was a correlation between an entry's quality and how often it was edited. In another case study, medical experts reviewed 35 *Wikipedia* articles on conjunctivitis, multiple sclerosis and otitis media with entries on similar topics from other popular online resources frequented by medical students⁹. The results found *Wikipedia* entries to be the easiest resource in which to find information. In addition, although *Wikipedia* entries were reasonably concise and current, they failed to cover key aspects of two of the topics and contained some factual errors. The report concluded that *Wikipedia* entries were thus unsuitable for medical students. Nevertheless, in a recent report published in *Psychological Medicine*, ten researchers from the University of Melbourne concluded that 'the quality of information on depression and schizophrenia on *Wikipedia* is generally as good as, or better than, that provided by centrally controlled websites, *Encyclopaedia Britannica* and a psychiatry textbook'¹⁰. For schizophrenia and depression, two commonly encountered psychiatric conditions, *Wikipedia* scored highest in the accuracy, timeliness and references categories – surpassing all other resources, including WebMD, NIMH, the Mayo Clinic and *Britannica Online*.

In one study, among the humanities and the social sciences, *Wikipedia* was not found to be a reliable source of historical articles, with an overall accuracy rate of 80% compared to 95-96% among the other sources, which included *Encyclopaedia Britannica*, *The Dictionary of American History* and *American National Biography Online*¹¹. *Wikipedia's* performance in articles on Philosophy was found to be mixed in one study, with high rates of coverage and accuracy but high rates of omissions as well¹². In an impressive review of thousands of *Wikipedia* articles in political science, about every major party gubernatorial candidate who ran between 1998 and 2008, the author found that *Wikipedia* was almost always accurate when relevant articles on the topic existed¹³. The coverage of topics was often very good especially for recent or prominent topics, but not as good on older topics. Omissions were, however, found to be frequent.

Prior to *Nature's* seminal study in 2005, *Wikipedia* assessed the quality of its entries through its 'featured article' and 'good article' peer review process¹⁴, and more recently through an ongoing pilot study to collect feedback¹⁵, which involves readers and editors rating articles according to trustworthiness, neutrality, completeness and readability, as well as rating

⁷ Clauson KA, Polen HH, Kamel Boulos MN, Joan H Dzenowagis JH. Scope, Completeness, and Accuracy of Drug Information in Wikipedia. *Ann. Pharmacother.* December 2008 vol. 42 no. 12 1814-1821

⁸ Devgan L, Powe N, Blakey B, Makary M. Wiki-Surgery? Internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons* 205:3, September 2007, Pages S76-S77

⁹ Pender M, Lasserre L, Kruesi L, Del Mar C, and Anaradha S. 2008. Putting Wikipedia to the Test: A Case Study. Paper presented at to the Special Libraries Association Annual Conference, Seattle, June 16.

¹⁰ Reavley NJ, Mackinnon AJ, Morgan AJ, Alvarez-Jimenez M, Hetrick SE, Killackey E, Nelson B, Purcell R, Yap MBH and Jorm AF. Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. *Psychological Medicine*, Available on CJO 2011 doi:10.1017/S003329171100287X

¹¹ Rector LH. 2008. "Comparison of *Wikipedia* and Other Encyclopaedias for Accuracy, Breadth, and Depth in Historical Articles." *Reference Services Review* 36 (1): 7-22.

¹² Bragues G. 2007. "Wiki-Philosophizing in a Marketplace of Ideas: Evaluating Wikipedia's Entries on Seven Great Minds. Working paper. <http://ssrn.com/abstract/978177>.

¹³ Brown A. Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage. *World Politics* 63:1, 2011.

¹⁴ Wikipedia (2011) *Featured articles*, [Online], Available at http://en.wikipedia.org/wiki/Wikipedia:Featured_articles [Accessed 11/03/11].

¹⁵ Wikipedia (2011) *Article feedback*, [Online], Available at http://www.mediawiki.org/wiki/Article_feedback [Accessed 01/07/11].

their self-perceived qualification to comment. *Wikipedia* has continued to develop and refine its quality review processes in part as a result of the findings of the *Nature* study and of other similar studies. However, there has never been any attempt to replicate, better or extend *Nature's* study, across disciplines and languages. Such a study would not only allow a greater understanding of the accuracy and quality issues pertaining to *Wikipedia* entries but would also provide information on how such issues may be addressed and/ or resolved.

Recently, *Wikipedia's* founder Jimmy Wales highlighted the importance of such a task, i.e. a study inspired by the *Nature* study but employing greater rigour by carrying out the assessment of articles across languages and across a range of disciplines spanning the humanities and sciences, involving the following characteristics:

1. Assessments carried out by academics and scholars.
2. Assessments on each pair of articles carried out by multiple expert reviewers to establish inter-rater reliability and eliminate biases.
3. Reviewers to be blind to the source of the article.
4. A variety of constructs and dimensions relating to the quality, accuracy, style, references and overall judgment.
5. Using both quantitative and qualitative rating techniques.

The importance of such a study would lie in the examination of articles in more than just the English language and in subjects other than solely science. This would allow differences in levels of accuracy and quality across languages and subject domains to be identified, which would inform decisions in the future, e.g. for editor recruitment efforts and the design of expert feedback mechanisms.

The size, scope and complexity of undertaking such a study would require considerable preliminary information on the methodology and design, compilation and functioning of rating scales, recruitment and location of the experts, and analysis and interpretation of results. As such it was decided that prior to the commencement of such a study, a small-scale preliminary project drawing on empirical evidence would be essential to determine a sound research methodology, which is the reason that the present study was undertaken.

This pilot study has therefore been carried out to collect and review preliminary evidence to inform the design of a larger, future study. The intention is that the results of this preliminary report will establish the best possible research approach, begin to hypothesise the best way for *Wikipedia* to measure and communicate the accuracy and quality of articles and provide a well-founded justification for seeking funding for a comprehensive study. This pilot study has been carried out for the *Wikimedia Foundation* by Epic, in partnership with the Department of Education at the University of Oxford, UK. The methodology, analysis and results of the study are presented in this report, followed by a discussion of the findings and the conclusion of the report.

2. Aims and Objectives

Aims:

The aims of this pilot study are as follows:

1. To explore the opinions of expert reviewers regarding attributes relating to the accuracy, quality and style of a sample of Wikipedia entries across a range of languages and disciplines.
2. To compare the accuracy, quality, style, references and judgment of *Wikipedia* entries as rated by experts to analogous entries from popular online alternative encyclopaedias in the same language.
3. To explore the viability of the methods used in respect of the first two aims for a possible future study on a larger scale.

Objectives:

Research objective 1: To explore the opinions of expert reviewers pertaining to the accuracy, quality, references, style and overall judgment of *Wikipedia* entries.

Research objective 2: To compare the accuracy, quality, references, style and overall judgment of *Wikipedia* entries to those of alternative online encyclopaedias.

Research objective 3: To compare the accuracy, quality, references, style and overall judgment of *Wikipedia* entries with those of alternative online encyclopaedias in each language, i.e. English, Spanish and Arabic.

Research objective 4: To compare the accuracy, quality, references, style and overall judgment of *Wikipedia* entries with those of alternative online encyclopaedias in each academic discipline i.e. Humanities; Social Sciences; Mathematics, Physics and Life Sciences; and Medical Sciences.

Research Objective 5: To comment on issues of importance pertaining to the design and methodology in carrying out the study.

3. Research Methodology

Figure 3.1 below depicts the research methodology employed in the study. In summary, this consisted of 31 experts (academics and doctoral students) reviewing two pairs of articles each in their area of expertise and in their native language. The languages selected for the purpose of this study were English, Spanish and Arabic. The rationale for selecting the same is mentioned in section 3.1 below. The academic areas of expertise selected for the purpose of this study were (a) Humanities (b) Social Sciences (c) Mathematics, Physics and Life Sciences and (d) Medical Sciences. The rationale for selecting these four academic areas to classify both articles and the reviewers' areas of expertise, was that they correspond with the four main academic divisions at the University of Oxford, which is where this study was carried out. Further details on each aspect of the methodology are described in the sections that follow.



Fig. 3.1 Flowchart of research methodology.

3.1 Selection Criteria

3.1.1 Selection of Languages

As of July 2012, there were 285 different language versions of *Wikipedia*¹⁶. Three of the most popular world languages were included for the purpose of this study, based firstly on their popularity in terms of numbers of native speakers¹⁷ and secondly in terms of numbers of *Wikipedia* articles⁹, with the intention of choosing those with potential for a wide reach.

The top five world languages in order by numbers of native speakers were found to be Mandarin (Standard Chinese), Spanish, English, Hindi-Urdu and Arabic. These appear in the list of number of articles per language version of *Wikipedia* ordered as follows: English, Spanish, Chinese, Arabic and then Hindi-Urdu. The Chinese *Wikipedia* was found to be

¹⁶ Wikipedia (2012) *Lists of Wikipedias*, [Online], Available at: http://meta.wikimedia.org/wiki/List_of_Wikipedias [Accessed 12/07/12].

¹⁷ Wikipedia (2011) *List of languages by number of native speakers*, [Online], Available at: http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers [Accessed 16/04/11].

heavily censored and was therefore excluded as it would possibly confound the research results¹⁸. The three languages selected at the end of this process were:

1. **English:** The *de facto* language in the UK, Australia, USA, UAE and Malaysia and the unifying language for countries such as Bangladesh, Botswana, India, Hong Kong, Pakistan, Philippines and Tanzania.
2. **Spanish:** The official language of Spain, as well as the *de facto* or *de jure* language of a large number of countries in Latin America, among them: Mexico, Argentina, Bolivia, Chile, Colombia, Ecuador, Paraguay and Venezuela. In addition, Spanish is the predominant language in Equatorial Guinea, Africa.
3. **Arabic:** The official language of a large number of countries across the Middle East and North Africa, among them: Bahrain, Egypt, Kuwait, Oman, Qatar, Saudi Arabia, Algeria and Tunisia. Modern Standard Arabic is based on Classical Arabic and is the literary language used in most current, printed Arabic publications and spoken by the Arabic media.

These languages offer a range of numbers of total articles and average edits per article for *Wikipedia*, as shown in Table 3.1 below:

Language	Ranking for total number of Wikipedia articles	Total number of articles	Average number of edits per article (2.s.f)
English	1 st	4,003,764	136
Spanish	7 th	904,461	68
Arabic	25 th	186,414	58

*Table 3.1 Characteristics of Wikipedia articles in each of the three study languages.*¹⁹

3.1.2 Selection of Comparison Encyclopaedias in Each of the Languages

The criteria for the selection of the comparison encyclopaedia in each of the three languages were as follows:

Essential Criteria:

1. The encyclopaedia should be available online.
2. The encyclopaedia should be a popular choice among the native speakers of that language.
3. The encyclopaedia should cover a broad range of articles within each specific discipline.

¹⁸ Wikipedia (2010) *Task force/China*, [Online], Available at http://strategy.wikimedia.org/wiki/China_Task_Force [Accessed 01/07/11].

¹⁹ Wikipedia (2012) *Lists of Wikipedias*, [Online], Available at: http://meta.wikimedia.org/wiki/List_of_Wikipedias [Accessed 12/07/12].

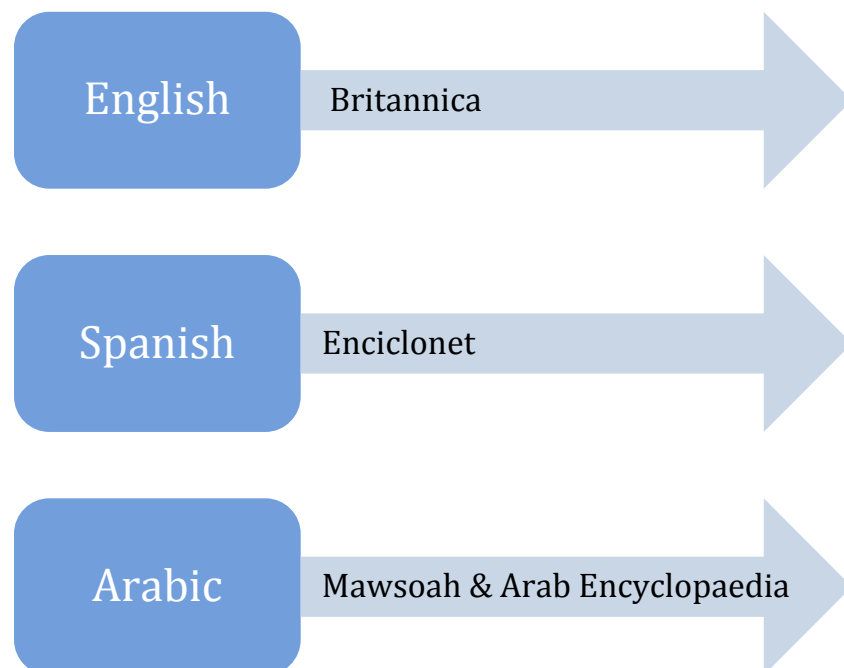
4. The encyclopaedia should contain articles of reasonable length on each of the topics selected as per the reviewers' academic area of expertise, i.e. at least 1.5 pages in length or more.

Preferable Criteria:

1. The encyclopaedia's articles should seem complete when read through by a native speaker of the language.
2. The encyclopaedia's articles should contain links at the bottom of the articles to enable the user to access further information if required.

The selection process of the encyclopaedia was based on the availability, quality and length of its articles. The selection was carried out by the research team with reference in each case to a native speaker in each of the three study languages (postgraduate students at the University of Oxford). The selection of comparative encyclopaedias for the study was made independent of the opinions of the research team at the Wikimedia Foundation. This was done in order to increase the robustness of the study design by eliminating any potential biases in the selection of the alternative encyclopaedias for comparison.

The following encyclopaedias were selected:



Encyclopaedia Britannica:

For the English language, the alternative encyclopaedia selected was the online home version of *Encyclopaedia Britannica*. As well as being the oldest English-language encyclopaedia, it was also the encyclopaedia originally chosen by *Nature* to compare with *Wikipedia*²⁰. *Britannica* was founded in 1768, in Edinburgh, Scotland, and has grown continuously since then with offices in London, New Delhi, Paris, Seoul, Sydney, Taipei and

²⁰ Giles, J. (2005) 'Internet encyclopaedias go head to head', *Nature*, vol.438, 15 December 2005, pp. 900-901.

Tokyo. The ownership of *Britannica* passed to two Americans in the 1930s and, since then, the company's headquarters has been in Chicago. *Britannica* was an early leader in digital publishing. In 1981, the first digital version of the *Encyclopædia Britannica* was created for the Lexis-Nexis service. It has been stated to be possibly the first digital encyclopaedia in the world. As personal computers grew in number in the mid-1980s *Britannica* produced the first multimedia CD ROM encyclopaedia in 1989. In 1994, *Britannica Online*, the first encyclopaedia on the Internet, was introduced²¹.

Enciclonet:

Enciclonet was selected to be the alternative encyclopaedia of choice in Spanish. *Enciclonet* is an online project based on the *Universal Encyclopaedia* and developed by Micronet equipment. It is described as the first online general encyclopaedia in Spanish (www.enciclonet.com). It was selected because of its high popularity, its high Alexa traffic rank of 322,628²² and because of the comprehensive nature of its articles. The other online Spanish encyclopaedias considered were *Enciclopedia Universal en Español* (which was not chosen as it could not be accessed in January 2012), *Ateneo de Cordoba*²³ (which was not chosen as it incorporated *Wikipedia* articles), *Gran Enciclopedia Aragonesa*²⁴ (which was not chosen because it was not found to be as comprehensive as *Enciclonet*), a little-known encyclopaedia developed by the University of Sevilla²⁵ and *Gran Enciclopedia de Espana*²⁶ (which was not found to be as comprehensive as *Enciclonet*).

Mawsoah and Arab Encyclopaedia:

*Mawsoah*²⁷ was selected to be the alternative encyclopaedia of choice in Arabic for the social sciences and medical sciences. *Arab Encyclopaedia*²⁸ was selected as the alternative Arabic encyclopaedia for mathematics, physics and life sciences. Due to extreme difficulty encountered in finding an online Arabic encyclopaedia to meet all four essential criteria, it was decided to select the best encyclopaedia choices for each academic discipline as there appeared to be a substantial segregation of encyclopaedias by discipline.

Mawsoah was selected because it has 150,000 articles and its articles appear to be comprehensive and have good categorisation. *Arab Encyclopaedia* was chosen because it appeared to have the highest traffic amongst the other alternative online encyclopaedias and has hyperlinks embedded into articles. Unlike *Mawsoah*, however, *Arab Encyclopaedia*'s articles are authored by a single person. In addition, it is extremely important to highlight that neither *Mawsoah* nor *Arab Encyclopaedia* covered all academic disciplines to the same extent, even for basic articles and articles on key concepts.

The other option considered for Arabic encyclopaedias was *Dahsha*²⁹, a Saudi Arabian encyclopaedia with high traffic. However, on exploring this option further, *Dahsha* did not appear to have the same coverage of topics as either *Mawsoah* or *Arab Encyclopaedia*.

²¹ Taken from http://corporate.britannica.com/company_info.html

²² <http://www.checkpagestats.com/www/enciclonet.com>

²³ <http://ateneodecordoba.org/index.php/Portada>

²⁴ <http://www.encyclopedia-aragonesa.com/>

²⁵ <http://www.us.es/>

²⁶ http://www.mienciclo.es/gee/index.php/Portada_GEE

²⁷ <http://www.mawsoah.net>

²⁸ <http://www.arab-ency.com>

²⁹ <http://www.dahsha.com/>

3.2 Sampling

3.2.1 Sampling of Expert Reviewers

Step 1: Selection of student reviewers

Student reviewers were recruited from the University of Oxford. All were postgraduate students, either currently studying or recently having completed either a masters or doctoral degree. 116 students were initially identified as potential reviewers, in order to cover the full range of academic disciplines and native languages selected for the study, 12 of whom were finally invited to participate (a further 12 were identified as a back-up). Each selected student was asked to provide biographical information in terms of educational qualifications, area of expertise and current academic focus (see Appendix I (2)).

Step 2: Identification and recruitment of established academics

Student reviewers identified academic experts known to them in their own areas of academic expertise. Criteria for nomination were as follows:

Essential Criteria

1. Each academic expert must have a higher educational qualification, preferably a PhD.
2. The academic expert must have demonstrated their academic status by having a permanent post at a highly rated department within a well-established University.
3. The academic expert should have worked closely with the student and have overlapping areas of research interests.
4. The academic expert should be fluent in the student's native language.

Desirable Criteria:

1. The academics and student should share the same native language.
2. They should have a number of publications in peer-reviewed journals, or be a leading investigator on a large-scale, funded project.

Each student was asked to nominate three academic experts and to provide contact details and a brief biography for each of three nominees. The list of nominees was reviewed by the research team to ensure they were eligible for participation. In the rare cases where the academic did not have a PhD, students were asked to nominate another academic in their stead. The final list of nominated academic experts totalled 33, out of which number 22 accepted the invitation from the project team to participate.

Step 3: Completion of review using online feedback tool

Reviewers were asked to review articles in their native language and relating to their area of academic expertise using an online review tool specially designed for the purpose of the

study. Of each pair of articles, one article was a *Wikipedia* entry on the topic and the other was an article on the same topic from the alternative online encyclopaedia for that language. Reviewers were not aware of the source of the articles and were asked to make no efforts to identify the same. All cues as to the source of the article were eliminated before the students viewed the article. This was carried out during the standardisation and anonymisation process, the details of which are described in section 3.4.2. Reviewers were asked to comment on the quality, accuracy, citability and style of each of the articles as well as on their opinions about the readability of the article and whether the information contained in it was, to the best of their knowledge, up to date. They were also asked to compare both articles within a pair, listing the strengths and limitations of each. Both quantitative and qualitative data were collected and reviewers were asked to confirm that they had made no attempt to identify the source of the articles by completing a declaration at the end of the review. The various dimensions assessed by the online feedback tool developed for the review process are detailed in Section 3.4.1.

3.3 Selection of articles

The selection of reviewers with strong academic credentials was considered to be paramount in this study, and therefore only after they had been recruited was it appropriate to seek articles that matched their areas of expertise sufficiently well.

A list of keywords for possible articles was drawn up based on the information provided by the students about:

1. Their area of research and academic expertise.
2. The nominated academic's area of research and academic expertise.
3. Areas of overlap between the students' and academics' areas of research and expertise.

As it turned out, it was not always possible to select articles that mapped the students' and academics' areas of expertise exactly, as articles for these niche areas were not found to exist in many encyclopaedias or were found to be incomplete or of inadequate length. A second phase was then embarked on by the research team to select articles of substantial length (≥ 1.5 pages) that appeared most complete and comprehensive. This resulted in a list of possible articles that was much broader and less specialist than initially sought, and which did not map on to the niche aspects of the academic's expertise. Thus the selection of articles was constrained by two important factors: one, the need to find topics appropriate for the academics whom we were able to recruit to the project; secondly, that articles from different online encyclopaedias were of comparable substance and focus. (Such factors would need to be taken carefully into account when embarking on a future large-scale study, where the demands of finding large numbers of comparable articles are likely to be considerable.)

Nevertheless, the second phase allowed the compilation of the 22 pairs of articles for review, across three languages and four academic disciplines. The topics of the articles selected for review are listed in Table 3.2.

The selection criteria for articles listed in table 3.2 were as follows:

1. The topic must be related to the academic and research interest of all the reviewers of the article.
2. Availability of an article on the topic in both *Wikipedia* and the alternative encyclopaedia of choice.
3. Length of the article on the topic in both *Wikipedia* and the alternative encyclopaedia of choice must be ≥ 1.5 pages when pasted into a MS Word document.
4. No traces of vandalism in the article (the definition of vandalism is given in Section 3.4.2). Note: This criterion turned out to have no impact on the selection of articles for the present study.

	Humanities	Social Sciences	Mathematics, Physics and Life Sciences	Medical Sciences
ENGLISH	Saint Thomas Aquinas/ Thomas Aquinas Saint Anselm of Canterbury/ Anselm of Canterbury	Elementary/ primary education Preschool education	Mutation Antibiotic resistance	Attention Memory
SPANISH		Cambio Climatico Energia Renovable Evo Morales Hugo Chavez	Número racional Polinomio	Neurona Percepción
ARABIC		Middle East Egypt	Mathematical proof Algorithm	Parkinson's disease Pharmacokinetics

Table 3.2 Final list of articles for review in each of the three study languages.

3.4 The Review Process

3.4.1 Development of a Feedback Questionnaire to Assess Articles

A feedback questionnaire was constructed following a literature review of current tools available to assess the quality and accuracy of written text. The feedback questionnaire was developed by the team.

It consists of 23 items that assess four key dimensions for assessing the quality of articles as follows:

1. Intrinsic attributes of quality and accuracy
2. Temporal attributes
3. Style
4. Subjective opinions

A variety of more detailed constructs was assessed under each of these dimensions using a Likert-type (i.e. 1-5) rating scale (see Appendix I (3)). These are listed in Table 3.3. Both qualitative and quantitative information was collected for each dimension.

Reviewers commented on each article within a pair using this feedback tool i.e. per reviewer; four such assessments were conducted corresponding to each of the four articles.

In addition, reviewers completed a comparative questionnaire after reviewing each pair of articles, where they were asked to comment about the two articles in the pair in comparison to each other (see Appendix 1(4)).

Dimension	Construct	Aspects Assessed
Intrinsic attributes of quality and accuracy	Accuracy/ Validity	Presentation of correct information, factual inaccuracies, errors, misleading statements
	Breadth of references	The extent to which the information is well researched and cited
	Quality of references	The relevance and importance of the references
	Completeness	All aspects of the topic addressed, omission of key facts
	Conciseness	Length of the article compared to the information contained in the text, presence of repetition
	Coherence	Coherence between different sections of the text
	Relevance	Extent of relevance of the information to the topic, presence of digressions
	Neutrality	Unbiased and objective nature of the information; acknowledgement of controversies and/ or gaps in knowledge
Temporal attributes	Currency	Information is up to date based on the reviewer's knowledge
Style	Writing style	Use of clear and appropriate language; spelling and grammatical accuracy, use of punctuation.
	Clarity and organisation	Structure of the article, order in which information is presented, readability
	Inclusion of photographs, charts and tables	Inclusion of photographs, charts and tables and their contribution to an understanding of the text
Subjective opinions	Enjoyment	The extent to which the reviewer enjoyed reading the article
	Citability	The extent to which the reviewer would cite the article in (a) non-academic work (b) academic work
	Strengths	Key strengths of the article
	Flaws	Key limitations of the article

Table 3.3 Dimensions and constructs of article feedback questionnaire.

The key articles in previous literature that informed the design of the tool used in this study were as follows:

1. Information Quality Discussions in Wikipedia. Stvilia B., Twidale M. B., Gasser L. and Smith C., 2005
2. Assessing information Quality of A Community-Based Encyclopaedia. Stvilia B., Twidale M. B., Smith C and Gasser L., 2005
3. http://www.mediawiki.org/wiki/Article_feedback/UX_Research
4. http://en.wikipedia.org/wiki/Reliability_of_Wikipedia
5. Crawford, H. (2001). Encyclopedias. In: R. Bopp, L. C. Smith (Eds.), Reference and information services: an introduction (3 ed.). (pp. 433-459). Englewood, CO: Libraries Unlimited
6. Gasser, L., Stvilia, B. (2001). A new framework for information quality. Technical report ISRN UIUCLIS--2001/1+AMAS. Champaign, IL: University of Illinois at Urbana Champaign
7. Critical Appraisal Skills Programme (CASP) Making sense of evidence: 10 questions to help you make sense of qualitative research
8. <http://strategy.wikimedia.org/wiki/Quality/Quality>
9. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. Kittur A., Kraut R. E. Proceedings of the 2008 ACM conference on Computer supported cooperative work
10. Measuring article quality in wikipedia: models and evaluation. Hu M., Lim E., Sun A., Lauw H. W. and Vuong B. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management

3.4.2 Standardisation and Anonymisation Protocol

A standardisation and anonymisation protocol was drawn up to ensure that all cues as to the source of the articles were removed. This included the removal of particular formatting patterns such as the article tree at the beginning of *Wikipedia* articles, special in-text references and internal links and the names of the article's authors.

Fig. 3.4 summarises the steps in the standardisation and anonymisation process. All standardisation and anonymisation was conducted by three researchers native in English, Spanish and Arabic respectively who were not part of the review panel of the study.

Step 1: Reading of article to identify vandalism

After pasting the article into a MS Word document, standardisers were asked to read through the article to identify any vandalism (this was of particular importance for *Wikipedia* entries which are open to edition by any user). **Vandalism** was defined as any addition, removal or change of content in a *deliberate* attempt to compromise the integrity

of the article³⁰. Examples of typical vandalism are adding irrelevant obscenities and crude humour to a page, illegitimately blanking pages and inserting obvious nonsense into a page. No instances of vandalism were detected in any of the articles for the present study, either by standardisers or reviewers.

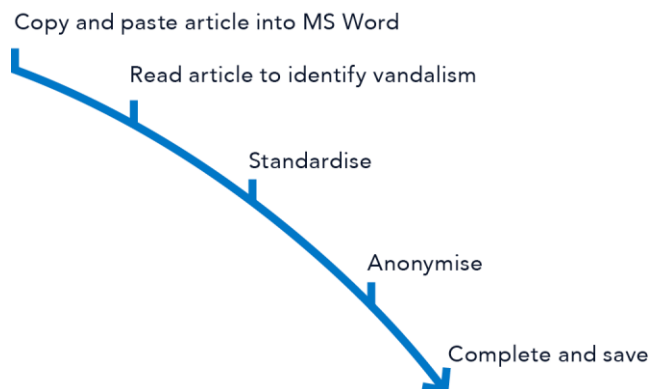


Figure 3.4 Summary of standardisation and anonymisation protocol.

Step 2: Standardisation Process

Article Text:

All articles selected from *Wikipedia* and from other popular online alternative encyclopaedias then underwent a process of standardisation to remove visible cues as to the source of the article. This included the conversion of all article text to black Arial font with specified font sizes for the title (16 Bold), Headings (14 Bold) and Sub-headings (10). All text was single spaced and aligned to the left.

Supporting Material:

Any supporting material e.g. photographs, flow-charts and plots was pasted at the end of the document section in which they appear, one after the other, in their order of appearance in the text. They were resized to 5cm x 5cm, and captions were pasted beneath the corresponding pictures in Arial font, size 10. In cases where the picture lacked a caption, one was not added.

References and Links:

References at the end of the text were maintained in a standard list format, in black Arial font (size 8). All hyperlinks from reference lists were removed and the presence of 'notes' at the end of *Wikipedia* entries were placed under references and formatted accordingly. All '**^ abcde**' were deleted from the references when they occurred.

For articles from alternative encyclopaedia choices, a heading entitled 'Additional Information (from links)' [Arial font, black, bold, size 14] was created at the bottom of the text of the primary article in the MS Word document. All articles under the assorted references sections were read through to confirm they are not covered in the text of the primary articles. Articles whose topics were not included in the primary articles were pasted

³⁰ <http://en.wikipedia.org/wiki/Wikipedia:Vandalism>

in the 'Additional Information (from links)' section in the order that they appeared in the primary article under sub-sections named after the title of the link and formatted as per the instructions mentioned in article text above. This procedure was not carried out for *Wikipedia* articles.

Step 3: Anonymisation Process

All articles then underwent a process of anonymisation to remove visible cues as to the source of the article. This included the following steps:

1. *Wikipedia* articles were read by the researchers to identify potential acts of vandalism as mentioned in Step 1.
2. Conversion to a standardised basic text format as mentioned in Step 2.
3. Removal of cues:
 - a) Certain characteristics cues such as the article tree in a *Wikipedia* entry, content warning (such as the 'article has multiple issues' box at the top of a *Wikipedia* entry), calls for donations, etc. were removed.
 - b) Block quotes in *Wikipedia* entries were formatted from italics to regular text in Arial font (Colour: black, Font Size: 10).
4. In text, references were maintained but hyperlinks, author names and affiliations were removed. The removal of authors' names was clearly essential in order to avoid making the origin of a particular obvious to reviewer, as indeed was the removal of the article tree from *Wikipedia* articles, because this information gave clear indications of the identity of encyclopaedias.
5. All 'See Also', 'Related Articles', 'External links', links to user ratings [*Wikipedia*], and 'links', 'related articles', 'share', 'like', 'get involved' features [*Britannica*], were removed.

An example of an article, standardised and anonymised according to this process and ready for review, is presented in Appendix I (6).

3.4.3 Development of the Online Review Tool

The articles and the article feedback questionnaires were uploaded onto an online review tool created using a Moodle. Moodle (www.moodle.org) is a Course Management System, also known as a Learning Management System or a Virtual Learning Environment. It is a free open source web application that educators can use to create effective online learning sites.

The objective of the online review tool was to:

1. provide an online platform for the experts to view, read and rate the pairs of articles accurately and easily and to make the review an enjoyable experience
2. facilitate easy collection of both quantitative and qualitative data for the purpose of data analysis

A username and password was generated for each reviewer enabling them to log into their account online and perform the following operations:

1. Consent to participate in the study.
2. Read the instructions for the review.
3. Access, view and read each article within a pair.
4. Comment on each article individually.
5. Comment on each article in comparison with each other.
6. Confirm that he/ she has completed the review himself/ herself and declare that he/ she has not made any attempt to identify the source of the articles.

4. Data Coding and Analysis

Fig. 4.1 depicts the processes relating to the coding of the data from the articles reviews, and the methods of quantitative and qualitative analysis employed.

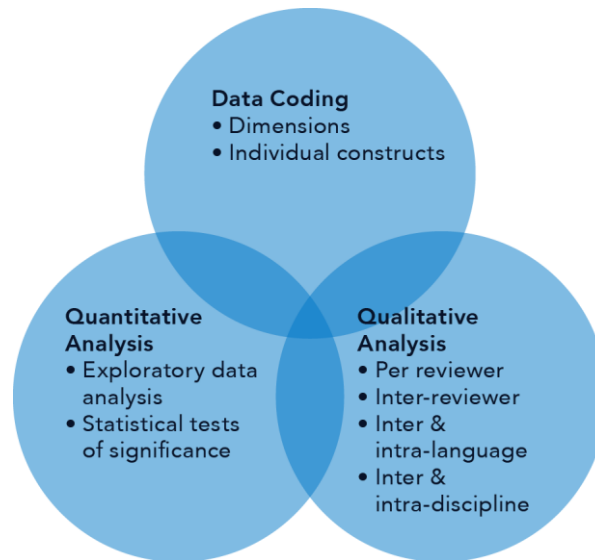


Fig. 4.1 Schematic depiction of the data coding and analysis process.

4.1 Data Coding

Data coding was carried out for the purpose of analysis and interpretation. The individual characteristics of each article commented upon by the reviewers (known as constructs) were collapsed into the five key dimensions as follows:

Accuracy:

This dimension represents the precision and correctness of the content of the article. It is computed by averaging the scores for validity, completeness, relevance, neutrality and currency.

References:

This represents the extent to which the article is adequately researched and referenced. It is calculated by averaging the scores for breadth and quality of references.

Style/ Readability:

Style/ readability represents the style and organisation of the article and the quality of the language, grammar, punctuation and visual aids used (if any). This dimension is computed by calculating the mean of the scores on conciseness, language, spelling and grammar, readability, enjoyment, clarity and organisation, coherence, photographs and pictures.

Overall Judgment:

This dimension represents the overall opinion of the reviewer and is computed by averaging the scores ranking the article's citability in an academic and non-academic piece of work. Citability was chosen to represent the reviewer's overall judgment of the article, as it was believed that a reviewer who considered an article to be of poor quality would be less likely to cite the article as compared to an article that he/ she considered to be of high quality. Citability was rated as cite worthy (1) and not cite worthy (0) and the score was averaged, thereby yielding a range from 0 to 1.

Overall Quality Score:

The overall quality score summarises the reviewer's opinion on the overall quality of the article. This is obtained by averaging the scores on the preceding four dimensions, i.e. accuracy, references, style/ readability and overall judgment.

Accuracy, references, style/ readability, overall judgment and overall quality scores were calculated per reviewer per article.

4.2 Quantitative Analysis

Fig. 4.2 depicts the stages in the quantitative analysis of the data. All quantitative data analysis was performed using the *Statistical Package for Social Sciences* version 15 licensed to the University of Oxford, UK. These various stages were carried out in order to explore the viability of arriving at findings about the overall spread of articles, and about distinct aspects of the articles (i.e. different languages and disciplines) that were specifically of interest within the study. The small scale of the present study does, it must be emphasised, mean that these detailed findings should be treated with some caution, but such tentative findings are valuable in indicating possible areas for future enquiry.

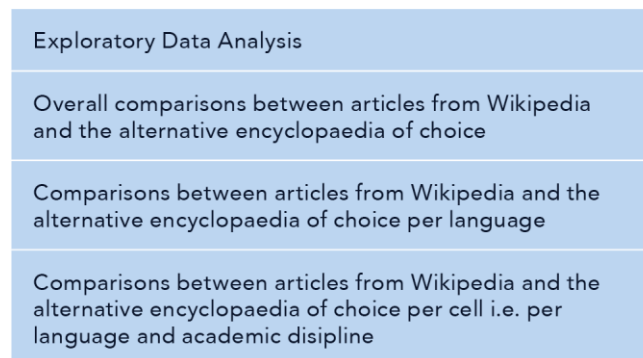


Fig. 4.2 Stages in quantitative data analysis.

4.3 Qualitative Analysis

Fig. 4.3 depicts the stages in qualitative analysis.

Blind analysis by subject	<ul style="list-style-type: none">• Identification of preferred articles from comments• Identification of issues common to multiple reviewers• Identification of criteria associated with highly positive comments
Key	<ul style="list-style-type: none">• Disclosure of key
Commonalities	<ul style="list-style-type: none">• Examination of commonalities and differences across the full sample of articles
Subject domains	<ul style="list-style-type: none">• Examination of commonalities and differences across subject domains
Languages	<ul style="list-style-type: none">• Examination of commonalities and differences across languages

Fig. 4.3 Stages in qualitative analysis.

The process of qualitative analysis followed the processes of reduction and display as recommended by Miles and Huberman, in their sourcebook on *Qualitative Data Analysis* (1994, Sage). Qualitative data were first of all summarised and compiled into spreadsheets for ease of comparison and analysis notes were written and revised over a period of time by reviewers in order to search for patterns, anomalies and illustrative examples. There was no question of using quantifiable content analysis on material such as this, given the fact that much of the language used had been generated by us in creating the criteria to be considered in the reviewer materials. Thus, it was the task of the qualitative data analysis to make interpretive judgments about salient themes and patterns, through repeated reading of the data followed by exploratory attempts at writing coherent and descriptions of results justifiable by substantial and wide-ranging use of illustrative material from the original raw data.

5. Results

The following section presents the results of this study. The results will be presented in two sub-sections based on the qualitative and quantitative analysis. The findings will be discussed with relationship to each other both in the context of this study and in the context of previous work in Section 6 (Discussion).

5.1 Quantitative Analysis

This section presents the findings following the quantitative analysis of the data from this study. The results of the quantitative analysis will be presented under the broad headings listed in Section 4.2 above.

Stage I: Exploratory Data Analysis

The characteristics of the dimensions for assessing the quality of articles in the entire sample are presented in table 5.1., and discussed in Section 3.4.1 above. The distributions of the dimensions are presented in table 5.2. Only the dimensions of accuracy and style/readability for the alternative encyclopaedia were found to be normally distributed. The remaining dimensions for both *Wikipedia* and the alternative encyclopaedia were found to be not normally distributed.

Dimension	Minimum	Maximum	Mean	Std. Deviation
Wikipedia (n=64)				
Accuracy	0.00	5.00	3.87	1.04
References	0.00	5.00	3.07	1.47
Style/ Readability	0.00	4.29	3.04	0.93
Overall Judgment	0.00	1.00	0.37	0.32
Overall Quality Score	0.65	3.26	2.18	0.65
Alternative Encyclopaedia (n=64)				
Accuracy	0.00	5.00	3.43	1.00
References	0.00	5.00	1.49	1.04
Style	0.00	5.00	3.47	0.99
Overall Judgment	0.00	1.00	0.33	0.37
Overall Quality Score	0.99	3.32	1.84	0.56

Table 5.1 Dimension Characteristics.

Dimension	Kolmogorov-Smirnov		Shapiro-Wilk	
	Statistic	Sig.	Statistic	Sig.
Wikipedia (n=64)				
Accuracy	0.16	0.00	0.89	0.00
References	0.16	0.00	0.89	0.00
Style/ Readability	0.14	0.00	0.93	0.00
Overall Judgment	0.12	0.04	0.96	0.05
Overall Quality Score	0.29	0.00	0.78	0.00
Alternative Encyclopaedia (n=64)				
Accuracy	0.08	0.20*	0.97	0.13
References	0.39	0.00	0.57	0.00
Style/ Readability	0.08	0.20*	0.97	0.13
Overall Judgment	0.13	0.01	0.94	0.00
Overall Quality Score	0.31	0.00	0.76	0.00

*p<0.05

Table 5.2 Dimension Distributions.

The sample characteristics in each of the languages and academic disciplines are presented in Table 5.3 and 5.4 respectively.

	Alternative Encyclopaedia		Wikipedia	
	Mean	SD	Mean	SD
English				
Accuracy	3.45	1.00	4.18	0.86
References	1.30	0.67	3.82	1.21
Style/ Readability	3.46	0.84	3.26	0.73
Overall Judgment	1.76	0.57	2.38	0.57
Overall Quality Score	0.30	0.45	0.32	0.29
Spanish				
Accuracy	3.46	0.90	4.00	0.95
References	1.52	1.10	3.40	1.23
Style/ Readability	3.56	0.92	3.11	0.93
Overall Judgment	1.85	0.54	2.27	0.69
Overall Quality Score	0.36	0.34	0.42	0.34
Arabic				
Accuracy	3.57	0.83	3.48	0.83
References	1.81	1.29	1.72	0.98
Style/ Readability	3.55	0.99	2.83	0.89
Overall Judgment	1.92	0.61	1.75	0.50
Overall Quality Score	0.34	0.30	0.38	0.34

Table 5.3 Sample characteristics according to language.

	Alternative Encyclopaedia		Wikipedia	
	Mean	SD	Mean	SD
Humanities				
Accuracy	3.85	0.57	4.30	0.66
References	1.38	0.75	4.13	0.75
Style/ Readability	3.96	0.63	3.43	0.35
Overall Judgment	2.09	0.50	2.57	0.34
Overall Quality Score	0.62	0.48	0.63	0.48
Social Sciences				
Accuracy	3.50	0.83	3.80	0.75
References	1.55	1.25	3.24	1.12
Style/ Readability	3.86	0.81	2.91	0.84
Overall Judgment	2.01	0.57	2.18	0.59
Overall Quality Score	0.58	0.34	0.47	0.39
Mathematics, Physics and Life Sciences				
Accuracy	4.28	0.48	4.24	0.99
References	1.97	1.26	3.19	1.72
Style/ Readability	3.58	0.89	3.42	0.78
Overall Judgment	2.07	0.57	2.31	0.62
Overall Quality Score	0.28	0.35	0.39	0.27
Medical Sciences				
Accuracy	2.77	0.70	3.71	1.00
References	1.14	0.30	2.77	1.45
Style/ Readability	3.10	0.89	2.99	0.97
Overall Judgment	1.45	0.33	2.00	0.72
Overall Quality Score	0.11	0.21	0.25	0.30

Table 5.4 Sample characteristics according to academic disciplines.

The sample characteristics of the entire sample categorised according to whether the articles were reviewed by student or academic experts are presented in Table 5.5.

	Student Experts		Academic Experts	
	Mean	SD	Mean	SD
Accuracy	3.64	0.94	3.74	0.94
References	2.16	1.34	2.40	1.53
Style/ Readability	3.15	1.00	3.80	0.85
Overall Judgment	1.96	0.63	2.03	0.63
Overall Quality Score	0.41	0.36	0.33	0.34

Table 5.5 Sample characteristics of entire sample according to nature of reviewer.

5.1.1 Overall Comparison across the Sample between Wikipedia Entries and Articles from the Alternative Encyclopaedias

The findings of the comparisons of reviewers' ratings of articles from *Wikipedia* and from the alternative encyclopaedias are presented in Table 5.6. *Wikipedia* articles were found to have scored significantly higher on the dimensions of accuracy, references, style/ readability and overall judgment (see Table 5.1).

	Test Statistic	P value
Accuracy	U= 2645.00**	0.004
References	U= 3285.00**	<0.001
Style/ Readability	U= 1533.00**	0.01
Overall Judgment	U= 2638.50**	0.001
Overall Quality Score	U= 2148.50	0.38

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.6 Comparison of article characteristics between Wikipedia and alternative encyclopaedias.

5.1.2 Comparison within each Language Group between Wikipedia Entries and Articles from the Alternative Encyclopaedias

The findings of the comparisons of reviewers' ratings of articles from *Wikipedia* and from the alternative encyclopaedias for English are presented in Table 5.7. Similar comparisons for Spanish and Arabic are presented in Tables 5.8 and 5.9 respectively.

In English, *Wikipedia* scored significantly higher on accuracy, references and overall judgment, as compared to the alternative encyclopaedia (*Encyclopaedia Britannica*) (see Tables 5.3 and 5.7). There were no differences between *Wikipedia* and *Encyclopaedia Britannica* on style and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 349.00*	0.01
References	U= 458.00**	<0.001
Style/ Readability	U= 222.00	0.64
Overall Judgment	U= 368.50**	0.003
Overall Quality Score	U= 272.50	0.43

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.7 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in English.

In Spanish, *Wikipedia* scored significantly higher on accuracy, references and overall judgment as compared to the alternative encyclopaedia (*Enciclonet*) (see Tables 5.3 and 5.8). There were no differences between *Wikipedia* and *Enciclonet* on style and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 459.00*	0.03
References	U= 578.00**	<0.001
Style/ Readability	U= 260.00	0.15
Overall Judgment	U=440.00*	0.01
Overall Quality Score	U= 342.00	0.53

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.8 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in Spanish.

In Arabic, the alternative encyclopaedias (*Mawsoah* and *ArabEncy*) scored significantly higher than on style than *Wikipedia* (see Tables 5.3 and 5.9). There were no differences between *Wikipedia* and either *Mawsoah* or *ArabEncy* on accuracy, references, overall judgment and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 130.00	0.96
References	U=133.50	0.84
Style/ Readability	U= 68.00*	0.02
Overall Judgment	U= 113.00	0.59
Overall Quality Score	U= 133.00	0.87

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.9 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in Arabic.

5.1.3 Comparison within each Academic Discipline between Wikipedia Entries and Articles from the Alternative Encyclopaedias

The findings of the comparisons of reviewers' ratings of articles from *Wikipedia* and from the alternative encyclopaedias for the Humanities are presented in Table 5.13. Similar comparisons for the Social Sciences, Mathematics, Physics and Life Sciences, and the Medical Sciences are presented in Tables 5.11, 5.12 and 5.13 respectively.

In the Humanities, *Wikipedia* scored significantly higher on references as compared to the alternative encyclopaedias (see Tables 5.4 and 5.10). There were no differences between *Wikipedia* and the alternative encyclopaedias on accuracy, style/ readability, overall judgment and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 12.00	0.34
References	U= 16.00*	0.03
Style/ Readability	U= 3.00	0.20
Overall Judgment	U= 12.50	0.20
Overall Quality Score	U= 6.00	0.69

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.10 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in the Humanities.

In the Social Sciences, *Wikipedia* scored significantly higher on references as compared to the alternative encyclopaedias, but the alternative encyclopaedias scored significantly higher on style/ readability as compared to *Wikipedia* (see Tables 5.4 and 5.11). There were no differences between *Wikipedia* and the alternative encyclopaedias on accuracy, overall judgment and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 238.00	0.30
References	U= 341.00**	<0.001
Style/ Readability	U= 95.00**	<0.004
Overall Judgment	U= 218.00	0.28
Overall Quality Score	U= 153.00	0.44

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.11 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in the Social Sciences.

In Mathematics, Physics and Life Sciences, *Wikipedia* scored significantly higher on references as compared to the alternative encyclopaedias (see Tables 5.4 and 5.12). There were no differences between *Wikipedia* and the alternative encyclopaedias on accuracy, style/ readability, overall judgment and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 191.00	0.37
References	U= 230.00*	0.03
Style/ Readability	U= 130.00	0.32
Overall Judgment	U= 206.00	0.17
Overall Quality Score	U= 198.00	0.27

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.12 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in Mathematics, Physics and Life Sciences.

In the Medical Sciences, *Wikipedia* scored significantly higher on accuracy, references and overall judgment as compared to the alternative encyclopaedias (see Tables 5.4 and 5.13). There were no differences between *Wikipedia* and the alternative encyclopaedias on style/ readability and overall quality score.

Dimensions	Test Statistic	P value
Accuracy	U= 384.00**	0.001
References	U= 409.00**	<0.001
Style/ Readability	U= 245.00	0.94
Overall Judgment	U= 359.50**	0.006
Overall Quality Score	U= 299.50	0.10

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.13 Comparison of article characteristics between Wikipedia and alternative encyclopaedias in the Medical Sciences.

5.1.4 Comparison between Wikipedia Entries and Articles from the Alternative Encyclopaedias per Cell i.e. per Language and Academic Discipline

The results of the intra-cell comparisons are presented in Table 5.14. It is difficult to interpret these findings without the raw data – a reporting of which is beyond the scope of this document. However, to summarise the findings, the far right column of the table contains the interpretation of the findings with reference to the database.

To summarise, in nine out of the ten cells, *Wikipedia* scored significantly higher than the alternative encyclopaedias on references. In one cell (English and Medical Sciences) *Wikipedia* scored significantly higher than the alternative (*Encyclopaedia Britannica*) on all dimensions. In another cell (Arabic and Mathematics, Physics and Life Sciences) the alternative scored significantly higher than *Wikipedia* on references, style and overall judgment.

	Accuracy	References	Style	Overall Judgment	Overall Quality Score	Interpretation Based on Raw Data
English & Humanities	U=12.00	U=16.00*	U=3.00	U=12.50	U=6.00	Wikipedia scored significantly higher than the alternative for references.
	P=0.34	P=0.03	P=0.20	P=0.20	P=0.69	
English & Social Sciences	U=6.00	U=16.00*	U=2.00	U=8.00	U=5.00	Wikipedia scored significantly higher than the alternative for references.
	P=0.69	P=0.03	P=0.11	P=1.00	P=0.49	
English & MPLS	U=26.50	U=35.5*	U=15.00	U=29.00	U=19.00	Wikipedia scored significantly higher than the alternative for references.
	P=0.18	P=0.002	P=0.70	P=0.09	P=1.00	
English & Medical Sciences	U=58.50**	U=59.00**	U=51.00*	U=57.00**	U=52.00*	Wikipedia scored significantly higher than the alternative on all dimensions.
	P=0.003	P=0.003	P=0.05	P=0.007	P=0.04	
Spanish & Social Sciences	U=94.50	U=112.50*	U=40.00	U=73.50	U=52.50	Wikipedia scored significantly higher than the alternative for references.
	P=0.20	P=0.02	P=0.07	P=0.40	P=0.61	
Spanish & MPLS	U=33.00*	U=36.00**	U=28.00	U=36.00**	U=33.00*	Wikipedia scored significantly higher on references overall judgment and overall quality score. The alternative scored higher on accuracy.
	P=0.02	P=0.002	P=0.13	P=0.002	P=0.02	
Spanish & Medical Sciences	U=41.00	U=55.00*	U=20.00	U=38.00	U=30.00	Wikipedia scored significantly higher than the alternative for references.
	P=0.38	P=0.02	P=0.23	P=0.57	P=0.88	
Arabic & Social Sciences	U=12.50	U=16.00*	U=3.00	U=15.00	U=10.00	Wikipedia scored significantly higher than the alternative for references.
	P=0.20	P=0.03	P=0.20	P=0.06	P=0.69	
Arabic & MPLS	U=7.50	U=3.00*	U=4.00*	U=4.00*	U=15.00	The alternative scored significantly higher on references, style and overall judgment.
	P=0.09	P=0.02	P=0.03	P=0.03	P=0.70	
Arabic & Medical Sciences	U=29.00	U=27.00	U=16.50	U=23.00	U=21.00	
	P=0.09	P=0.18	P=0.82	P=0.49	P=0.70	

*p<0.05, **p<0.01. U = Mann Whitney U test statistic. MPLS = Mathematics, Physics and Life Sciences.

Table 5.14 Intra-cell comparisons of articles between Wikipedia and alternative encyclopaedias.

5.1.5 Inter-Reviewer Comparisons

There were no differences in the scoring of articles for both *Wikipedia* articles and in articles from the alternative encyclopaedias, based on whether the articles were reviewed by students or academic experts. These results are presented in Table 5.15.

Dimensions	Wikipedia	Alternative Encyclopaedia
Accuracy	U=437.50, p=0.97	U=469.00, p=0.67
References	U=441.00, p=0.99	U=527.00, p=0.13
Style/ Readability	U=484.50, p=0.52	U=483.00, p=0.53
Overall Judgment	U=370.00, p=0.84	U=378.50, p=0.35
Overall Quality Score	U=443.50, p=0.32	U=493.50, p=0.41

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.15 Comparisons in ratings of articles (Wikipedia and alternative encyclopaedias) based on whether scored by student or academic experts.

The results of the inter-reviewer comparisons categorised according to language are presented in table 5.16 below. Spanish academics scored articles significantly higher for style/ readability and overall judgment as compared to students. Overall quality scores were significantly higher among students native in English, compared to academics native in English, although no significant differences were detected on any of the other four dimensions. A similar finding was found for overall quality scores among Arabic reviewers.

The results of the inter-reviewer comparisons categorised according to academic disciplines are presented in table 5.17 below. There were no significant differences between the ratings of articles by students and academic experts in the Humanities, Social Sciences, Mathematics, Physics and Life Sciences, and the Medical Sciences.

	Student Mean (SD)	Academic Mean (SD)	Statistical Test	p value
English				
Accuracy	3.32 (1.07)	3.92 (0.84)	U = 363.50	p = 0.13
References	1.97 (1.22)	2.69 (1.58)	U = 350.00	p = 0.21
Style/ Readability	2.75 (0.97)	3.61 (0.82)	U = 408.00*	p = 0.02
Overall Judgment	1.72 (0.61)	2.22 (0.61)	U = 391.00*	p = 0.01
Overall Quality Score	0.28 (0.31)	0.44 (0.34)	U = 340.00	p = 0.12
Spanish				
Accuracy	3.38 (0.36)	3.58 (0.93)	U = 117.50	p = 0.36
References	1.63 (0.92)	1.81 (1.21)	U = 102.00	p = 0.82
Style/ Readability	3.24 (1.08)	3.17 (0.99)	U = 91.00	p = 0.85
Overall Judgment	1.77 (0.39)	1.86 (0.60)	U = 104.00	p = 0.75
Overall Quality Score	0.31 (0.26)	0.38 (0.34)	U = 103.50	p = 0.75
Arabic				
Accuracy	4.10 (0.85)	3.66 (1.05)	U = 165.50	p = 0.15
References	2.63 (1.61)	2.52 (1.62)	U = 214.50	p = 0.81
Style/ Readability	3.50 (0.90)	3.28 (0.71)	U = 182.50	p = 0.31
Overall Judgment	2.28 (0.63)	1.95 (0.64)	U = 155.50	p = 0.10
Overall Quality Score	0.59 (0.37)	0.14 (0.27)	U = 83.50*	p <0.001

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.16 Comparisons in ratings of articles (Wikipedia and alternative encyclopaedias) by student or academic experts, categorised according to language.

	Student Mean (SD)	Academic Mean (SD)	Statistical Test	p value
Humanities				
Accuracy	3.85 (0.57)	4.30 (0.66)	U = 7.00	p = 0.89
References	1.38 (0.75)	4.13 (0.75)	U = 8.00	p = 1.00
Style/ Readability	3.96 (0.63)	3.43 (0.35)	U = 7.00	p = 0.89
Overall Judgment	2.09 (0.50)	2.57 (0.34)	U = 6.50	p = 0.69
Overall Quality Score	0.63 (0.48)	0.50 (0.19)	U = 6.50	p = 0.69
Social Sciences				
Accuracy	3.42 (0.86)	3.82 (0.72)	U = 230.00	p = 0.29
References	1.86 (1.02)	2.77 (1.61)	U = 232.00	p = 0.26
Style/ Readability	3.29 (1.08)	3.45 (0.85)	U = 188.50	p = 0.92
Overall Judgment	1.92 (0.47)	2.23 (0.63)	U = 226.50	p = 0.14
Overall Quality Score	0.50 (0.32)	0.54 (0.41)	U = 189.00	p = 0.72
Mathematics, Physics & Life Sciences				
Accuracy	4.40 (0.45)	4.22 (0.84)	U = 106.00	p = 0.84
References	2.94 (1.47)	2.48 (1.66)	U = 94.50	p = 0.51
Style/ Readability	3.27 (1.09)	3.51 (0.76)	U = 125.00	p = 0.64
Overall Judgment	2.35 (0.69)	2.15 (0.58)	U = 92.00	p = 0.47
Overall Quality Score	0.56 (0.42)	0.27 (0.25)	U = 65.50	p = 0.08
Medical Sciences				
Accuracy	3.30 (1.07)	3.22 (0.96)	U = 173.50	p = 0.62
References	1.83 (1.39)	2.00 (1.33)	U = 218.00	p = 0.46
Style/ Readability	2.70 (0.89)	3.17 (0.92)	U = 248.50	p = 0.14
Overall Judgment	1.62 (0.66)	1.76 (0.61)	U = 220.00	p = 0.46
Overall Quality Score	0.13 (0.23)	0.20 (0.28)	U = 217.50	p = 0.42

*p<0.05, **p<0.01. U = Mann Whitney U test statistic.

Table 5.17 Comparisons in ratings of articles (Wikipedia and alternative encyclopaedias) by student or academic experts, categorised according to academic discipline.

5.2 Qualitative Findings

This section of the report summarises and discusses findings from the qualitative element of the research, in terms of the perceptions, opinions and judgments of the expert reviewers regarding the articles from *Wikipedia*, and other online encyclopaedias. In Section 3, we explained how reviewers – both professional academics, and graduate students – were asked to comment on the quality, accuracy, citability and style of a few articles each, in their own fields of expertise. As was shown in that section, we paired *Wikipedia* articles with similar ones from the following sources: online *Encyclopaedia Britannica* (English articles), *Enciclonet* (Spanish), *Mawsoah* and *Arab Encyclopaedia* (Arabic), removing all evidence of the source of each article. We asked reviewers to comment on a range of quality criteria, summarised as *accuracy*, incorporating validity, completeness, relevance, neutrality/ bias, currency; *use of references*; *style/ readability* incorporating conciseness, language, spelling and grammar, coherence, use of illustrative material, and enjoyment. Having commented on each of these aspects for each paper separately, reviewers were asked to compare the two ('Please use the space below to make any additional comments about the two articles in comparison with each other').

5.2.1 Academics' Qualitative Judgments

In this section we shall look first of all, in section 5.2.1.1, at how the academics in this study tended to make judgments, both positive and negative, about the full range of online encyclopaedia articles in the sample. Then we shall look specifically in 5.2.1.2 at the question of whether it is possible to identify strengths and weaknesses that are characteristic of *Wikipedia* articles in particular. Academics were simply told that the articles given to them for review 'have been carefully chosen from popular online encyclopaedias to overlap with your area of academic expertise', and were urged not to attempt to identify the origins of articles. Therefore we aim to detect whether judgments, made blind as to the identity of different articles, revealed characteristic patterns regarding *Wikipedia* articles.

Whilst different reviewers sometimes expressed contradictory opinions regarding the same article (which will be discussed further in Section 6), initial analysis has indicated that there was no marked pattern of differences of opinion between student reviewers and more established academic reviewers. For that reason, we include responses from both in the following sections of this report. We do, though, in the interests of transparency, indicate throughout the qualitative data whether a comment came from a student, an established academic or a professor. It should be noted that the term 'student' covered masters students, research students reading for a variety of research degrees including doctorates, and postdoctoral students. And again, we specify these distinctions when quoting comments from reviewers.

5.2.1.1 Academics' Judgments about Online Encyclopaedia Articles in General

In this sub-section, we try to capture the aspects of articles that earned either the approval or disapproval of the academic reviewers across the full sample of articles selected. The aspects praised and criticised below are fairly evenly distributed across all sources, and are intended to illustrate the kinds of judgments about online encyclopaedia articles across the sample that were generated by the experience of comparing and reviewing one or more pairs of articles on specific topics. So although we identify the sources of the articles in the following examples, again in the interest of transparency, we are not suggesting that the characteristics discussed in this sub-section are uniquely typical of that source.

It was evident on a number of occasions that academics, having considered the criteria put before them to help them with their quality judgments, and scored each of these criteria individually, very often went on to consider the combined characteristics of an article, balancing and synthesising the individual elements to arrive at an overall judgment. This is evident for example in the responses of two different reviewers for the pair of articles on *Evo Morales*, in which it appears that the overall feel and coherence of the article were valued more highly than quantity or currency of information:

- "The second one [*Enciclonet*] is much better than the first one [*Wikipedia*]. It is shorter but contains all the important information up to 2006. It also has a better and more professional style." (Reviewer 3 – academic)
- "Speaking as an academic, I much preferred the second piece [*Enciclonet*] in this case. However, this was on principally intellectual and aesthetic grounds. For the most up to date information, and for more information about the various critics of

Morales' governance, I would have to consult the first piece [*Wikipedia*] since this is absent in the second.” (Reviewer 2 – academic)

For a number of academics, the impression of the article as a cohesive piece of writing appeared to be valued at least as much as the extent of subject matter, as can be seen in this comment on the pair of articles on *Energia Renovable*:

- “The first article [*Wikipedia*] is more suitable and better descriptive of the subject matter. However, article 2 (*Enciclonet*) is half its size and can be read more quickly, which poses distinct advantages. Both are well written and accurate.” (Reviewer 1 – Professor)

Even when lack of comprehensiveness (specifically here, lack of currency) is acknowledged as undermining its usefulness, the overall feel of an article still earns it some degree of approval, even if not actually rendering it preferable to the article that is more up to date:

- “The first article [*Wikipedia*] had more information, but the second [*Mawsoah*] was much more eloquent.” (Reviewer 1 – research student – on articles on *Egypt*)

By the same measure, when one article abandoned any attempt at providing an engaging narrative, it was viewed quite negatively:

- “There is not a lot of writing in this article [*Wikipedia*]; mostly there are series of factual information in bullet point format.” (Reviewer 2 – academic – *Primary Education*)

Criticisms of this kind were made about both *Wikipedia* and non-*Wikipedia* articles, and are quoted here in order to capture the impression emerging from the data that academics – while of course strongly concerned with accuracy, currency and comprehensiveness (also demonstrated in the quantitative results) – also judge articles of this kind for their capacity to bring a topic alive to the non-expert or casual reader. This theme runs through the comments from all three reviewers on the pair of articles about *Polonomia*:

- “The second article [*Wikipedia*] is much clearer and concise, though it could need some additional information. The first article [*Enciclonet*], on the other hand, lacks focus and is rather inconsistent.” (Reviewer 3 – academic)
- “The first article is too confusing, poorly written and makes emphasis on one aspect of the theory. The second one is better written, gives a good overview, but does not cover in depth any topic. The scope of the first is larger but the execution is very poor. The quality of the second is higher, but it is too short.” (Reviewer 2 – postdoc)

Certain tendencies, which to some extent crossed both language and disciplinary boundaries, are apparent here. Concision (which above all seems to have been taken to mean something along the lines of getting straight to the point) is valued a great deal by many reviewers (especially with respect to scientific articles) as is good writing – in terms of having a clear and informative tone of voice throughout:

- “Article 1 [*Enciclonet*] goes deeper than the article 2 [*Wikipedia*]. In particular, notions of polynomial in higher algebraic settings are discussed and more theoretical

results are cited. Also, it does a better job when discussing ‘factorización de polinomios’ (polynomial factorisation). On the other hand, article 2 [Wikipedia] is way better written than article 1. It has the right encyclopaedic tone, including a very good introduction. Also, article 2 does the very important task of pointing to applications of the subject at hand.” (Reviewer 1 – academic)

The following picks this up very clearly and reiterates the emphasis from many reviewers that for an encyclopaedia article to be ‘well-written’, crucially entails it being accessible to the kind of readership presumed to seek out online encyclopaedia articles:

- “This article [Britannica] is factually correct and gives some interesting historical background information on antibiotic resistance. The article is written in a style that is simple, and this article should be accessible to both specialist and non-specialist readers. The article avoids an excessive use of technical jargon, and instead focuses on ‘real’ world examples of antibiotic resistance. Good, logical structure.” (Reviewer 2 – academic – *Antibiotic Resistance*)

The same reviewer, discussing the Wikipedia article on the same topic, in fact recognised that this provided richer content:

- “The second article [Wikipedia] provided much more detailed information on antibiotics and resistance, including very good citations to the scientific literature. However, the [...] article lacked organisation and structure.” (Reviewer 2 – academic – *Antibiotic Resistance*)

Thus, values such as simplicity, accessibility, lack of jargon and good structure are very often emphasised alongside values more immediately associated with encyclopaedias, such as accuracy, comprehensiveness and currency of information. It seemed, generally, that the academics reviewing these articles were generally willing to accept certain deficiencies in online encyclopaedia articles, so long as they combined some degree of accuracy, currency and scope with an account that brought a subject to life for newcomers to an area. In that respect, substantial content was no substitute for the lack of an underlying dynamic or coherence in its account of the topic, which is clearly considered by these reviewers to be a problem with respect to each of the following articles:

- “It [Wikipedia] is a very shallow article [...] there are many terms used. I did not find any reference for, the only reference used is inappropriate [...] there is no biased info in the article, because there is no controversy in the article, it is just explanation of the medical terms [...] there is no coherence in the article because it is just stating terms and jumping from one term to another without any connection.” (Reviewer 1 – academic – on *Pharmacokinetics*)
- “The second article [Britannica] is way too long; it is not good enough to warrant such a long piece.” (Reviewer 2 – academic – on *Memory*)

In terms of articles judged as very poor, such as the two above, the factors that led to harsh judgments appeared to be common to all sources, and it is hard to locate anything specific to any encyclopaedia in such judgments. Strongly negative reviews of articles generally consisted of an accumulation of weak points in terms of accuracy, missing information,

weak structure and lack of clarity, unredeemed by any strong impression of usability or readability:

- “The text is too short and I don't think it is concise [...] The information is not well structured. In certain sentences we don't understand what the author is focused on [...] the use of the example is not in coherence with what the author intended to explain [...] It would have been useful to include a history section, and mention further topics such as interaction with logic and foundations of Mathematics [...] While most key ideas are included, no pointers are provided for the topics mentioned in the article, nor are there examples for them.” (Combination of comments from all three reviewers for *Wikipedia* article on *Mathematical Proof*)
- “The article is poorly written. Its language is stiff and it has a number of errors [...] The all-important relation of rational numbers (‘números racionales’) with real numbers is omitted completely. Indeed, there is no mention of real numbers at all [...] The grammatical and factual errors and the dull tone preclude the article from being useful in this regard [...] The article is repetitive and extremely boring. Moreover, it is pretentious to spend seven pages discussing such basic ideas [...] Would just confuse a non-academic reader. Too elementary [...] The article needs to be re-written, possibly from scratch. It has absolutely no value.” (Combination of comments from all three reviewers for *Enciclonet* article on *Numero Racional*)

Overall, quite a small number of such articles were identified, whether from *Wikipedia* or other sources. In addition the article on *Mathematical Proof* quoted above, largely negative judgments about *Wikipedia* articles applied just to a small number out of the total of 22 *Wikipedia* articles: *Pharmacokinetics*, *Percepción* (from the Spanish version), *Primary Education* and (according to just one of the two reviewers) *St Thomas Aquinas*. A certain number of negative comments were made about most articles, because the academic reviewers were generally rigorous in their judgments, but these were usually balanced or redeemed by a very fair identification of strengths. In the next section we focus mainly on *Wikipedia* articles in order to explore the balance of qualities that was identified in these specifically, and in order to see if it is possible to detect a particular mix of strengths and weaknesses that are particularly relevant to *Wikipedia*.

5.2.1.2 Academics' Judgments about *Wikipedia* Articles in Particular

When it comes to articles that were judged as being satisfactory or good, which is to say an article that is readable, and provides a useful point of reference or a good introduction to a topic, we argue that it was indeed possible to detect a particular pattern of qualities that were particularly characteristic of the *Wikipedia* articles within this sample at least. Not all of these qualities are wholly positive if taken on their own, but nonetheless constitute a set of characteristics which in combination outweigh specific weaknesses. The *Wikipedia* article on *Hugo Chavez*, for instance, illustrates this particular combination of qualities: “Generally speaking, the second article [*Wikipedia*] was much stronger than the first. It was far more comprehensive and detailed, it was up to date (going right up to the middle of last year rather than more or less stopping in 2005), and it offered a far more politically neutral interpretation of the subject.” Both reviewers agreed that the *Wikipedia* article was the stronger on this particular topic, despite the fact that it did possess certain weaknesses:

- “The only areas it fell down on were length (the second being three times as long) and [...] lacking a clear argument or unifying perspective about the subject, making it a little harder to isolate what the key issues for debate might be.” (Reviewer 1 – academic)

In general, it appears that *Wikipedia* articles were distinguishable from other online encyclopaedias in the qualitative judgments with respect to the following characteristics if combined within a particular article: good *coverage of topic*, *currency*, *quality of referencing*, along with the less desirable qualities of *redundancy and repetition*. We would add to this also the non-appearance of a particular area of potential criticism – that of *bias*, as this did not appear to be an issue on more than a very few occasions in judgments made about *Wikipedia* articles.

Coverage of topic

- “The second article [*Mawsoah*] can be part of the first one [*Wikipedia*] [...] the first article is comprehensive while the second one is just an introduction, so we can use some information in the second article which is missing in the first one, like the addicts story to complete the first one.” (Reviewer 3 – academic – *Parkinson’s*)

In the same spirit, all three reviewers for the article on *Memory* felt that, despite “minor flaws” (Reviewer 2), the *Wikipedia* article was superior especially with respect to its coverage of the topic:

- “The first article [*Wikipedia*] is decent. It is reasonably concise, and covers most things that I would include – certainly it is not perfect, and there are things missing, but it is concise and well-written. By contrast the second article [*Britannica*] is very vague and makes minimal links to the actual original science behind the points [...] I actually think that it would be a little misleading to a novice, because the literature has developed so much in the last 10-15 years.” (Reviewer 3 – doctoral student)

Currency

Comprehensive coverage tended to be taken to imply currency as well, and this was certainly an area where *Wikipedia* articles consistently scored higher than others in the qualitative judgements:

- “I think that the real strength of this article is that it gives people a good overview of what *Attention* actually is. It covers the historical background of the research area, but also more up to date perspectives. It is also very transparent about the overlap between attention and other related areas of study, such as Working Memory and Executive Processes [...]”

Indeed, this article earns particular high praise from this reviewer specifically because of its currency:

- “Everything in this article [*Wikipedia*] is stuff that I would have included had I written it myself. It is also very 'current' – all the stuff on disorders of attention in children is really very new [...] Everything that is stated as fact is pretty much accepted by the

majority of the literature. I cannot really see any particular perspective coming through here. It is actually very carefully written.” (Reviewer 1 – academic – *Attention*)

This judgment contrasts sharply with the comments from the same reviewer on the *Britannica* article on the same topic, which was described as “all very out-of-date, and therefore would be of no use in a current research article”. In fact, it was consistently striking throughout the sample, that *Wikipedia* articles were nearly all considered to be more up to date than others, although being more up to date was not always a sufficient reason for a *Wikipedia* article to be considered the better of the two, if it failed on a combination of other factors such as clarity, cohesion or accuracy.

Referencing

The same is also true with respect to the presence of references. *Wikipedia* articles generally earned approval for their references, although – as we indicate below – these were not invariably judged to be advantageous. *Wikipedia* articles were clearly acknowledged as being more extensively referenced than others, even if in some respects it was not considered to provide as much information as the other article:

- “The second article [*Wikipedia*] gives a clear idea of the nature of climate change and its science but doesn't give as much detail on its impacts as the first one. Overall, the second article is better structured, organised and referenced,” (Reviewer 2 – doctoral student – *Cambio Climatico*)

When the references are considered to be appropriate, this would generally earn the highest praise from reviewers, as in the following comment from the other reviewer for this article: “References are broad, valid and of the highest quality available.” (Reviewer 1 – professor – *Cambio Climatico*). Similarly, it is the scholarly nature of its writing, supported by appropriate references, that earns *Wikipedia* higher praise in the following two separate instances, even though it is by implication to some extent insufficiently comprehensive if taken alone:

- “I preferred the 1st article [*Wikipedia*] to the second. It is written in a more scholarly manner and it provides a lot of references. I found the second paper still a draft, and this might be the case. Ideally you would combine the two to give a more comprehensive picture of preschool education.” Reviewer 2 – academic – *Preschool Education*)
- “The works cited are all of high-scholarly quality.” (Reviewer 1 – research student – *Anselm of Canterbury*)

The mere existence of references did not, though, necessarily earn approval, as all three reviewers make quite clear with respect to the *Wikipedia* article on *Parkinson's*:

- “The references cited are all from internet and these references could be changed or removed from the internet [...] I prefer the use of medical data from published text books in the right way.” (Reviewer 1 – academic)
- “Referencing was poor throughout the article.” (Reviewer 2 – academic)

- “References used are internet websites, no journals or books are used.” (Reviewer 3 – student)

A similar point is made about the *Wikipedia* article on *Mutation*:

- “Many of the references are from websites, magazines, or other popular media, and not from primary source scientific articles.” (Reviewer 2 – academic)

But the same reviewer pointed out, in discussing the fact that the *Britannica* article “mentions the mutation rate of HIV, but doesn't cite any HIV related material”, that anyway, “much of the article is quite basic and does not necessarily need intensive citation.” Thus the mere presence of references is not inevitably viewed as an advantage if (a) the references are of a generally low level and (b) the overall article appears to aspire to be simply a good basic introduction to a topic.

This reflects a more general feeling that a good article needs to balance its elements throughout:

- “All the references are published by recognized journals or are books written by academics that work in the topic. [...] The article [*Wikipedia*] is concise and focuses on its topic. All the information provided is relevant and necessary [...] provided in a well-structured form.” (Reviewer 4 – academic – *Neurona*)

but

“The use of technical terms is not accompanied by an explanation [...] it would be necessary to add new information to complete the map. [...] I would probably eliminate the topic about artificial neural networks.” (Reviewer 4 – academic – *Neurona*)

For most reviewers, though, lack of references was sometimes seen as a negative feature, regardless of other qualities, as the same reviewer makes clear with respect to the *Enciclonet* article on *Neurona*: “It is a great article, well-structured, clear, and easy to understand and read. The information provided is precise and complete. However, no references are provided and no topics are treated in depth.” (Reviewer 4 – academic)

Redundancy and Repetition

The following three comments on *Wikipedia* articles represent what was quite a common theme from many reviewers of *Wikipedia* in particular, which is to say up to date content with good coverage of issues, but at the same time a tendency to repetition and redundancy of content:

- “A lot of information, including a very thorough account of the events of Anselm's life. Mentions his most important ideas and works and discusses them reasonably well. Cites respectable scholarly sources, for the most part. Doesn't read completely smoothly, a bit repetitive at times. There are some digressions and random sentences that harm the overall coherence.” (Reviewer 1 – academic – *St Anselm*)

While being a bit repetitive does not necessarily constitute a serious problem, simply repeating the same information at some length is clearly seen as a potential source of distraction or potential irritation to the reader:

- “Within different sections, the article is [...] generally well structured, although there are some cases where information is repeated in multiple sections. For example, information on the applications of antibiotics in genetic engineering is given at the end of the article and in the section on mechanisms of antibiotic resistance; within the section on mechanisms of antibiotic resistance, general information is provided on mechanisms of antibiotic resistance followed by very specific information on mechanisms of resistance to one class of antibiotics (fluoroquinolones).” (Reviewer 1 – academic – *Antibiotic resistance*)

The point made by this reviewer of the *Wikipedia* article on *Evo Morales* makes an important general point which does appear to have been made considerably more often with respect to *Wikipedia*:

- “It feels a little disaggregated at times [...] The piece is fine but not exceptional in terms of overall coherence. It doesn't read as if someone has thought about the whole text as a reading experience, whether the author or editor. So it's fine if you're delving in to get a particular fact, but it doesn't work amazingly well as a singular read.” Reviewer 2 – academic – *Evo Morales*)

The notion of the ‘singular read’ frequently surfaces in one way or another in article reviews, as indicated in the previous sub-section of this report, and is one that should clearly be considered seriously, in terms of its impact upon reading experience. More serious still, perhaps, is the suggestion from both the other reviewers for the same article that internal inconsistencies had resulted in actual contradictions in the content:

- “I think that there is a very marked shift between the first ‘biographical’ part of the article, and the part that begins with the 2005 election victory of Evo. The first part is rather ‘Evo-friendly’ and relies almost exclusively on direct quotes from Evo. The second part of the article is rather more ‘anti-Evo’ and relies on newspaper references. It almost seems at times that it was written by two different people.” (Reviewer 1 – research student – *Evo Morales*)
- “Moves from statements that are too favourable to Morales to some statements that are too critical without enough support. Weak sourcing and bibliography. Would not cite in non-academic piece as not rigorous or well-organised enough.” (Reviewer 3 – academic – *Evo Morales*)

However, there were in fact very few indications of any significant degree of internal contradiction identified in the broad sample of *Wikipedia* articles. The problem identified here concerns a lower level, but nonetheless important, issue of a lack of consistency and cohesion arising from the multi-authorship of articles.

Neutrality/ Bias

The issue of bias did not often appear to arise from this sample as a major threat to the quality of articles from *Wikipedia*, or indeed any other sources. It was generally referred to

because the review process asked for a judgment on the question of bias, and reviewers were certainly careful to pay it due attention, even if occasionally they made it clear that the topic was not one, anyway, where issues of bias were likely to arise: “Not a very controversial topic” (Reviewer 3 – doctoral student – *Mutation*). Some references to bias suggest a slight modulation in the use of the term, such as in the review of *Numero Racional*, where the second reviewer considered that the *Enciclonet* article was “biased towards a formalist and algebraic point of view”. In this instance at least it seems that ‘bias’ is used to describe insufficient scope in the discussion of a particular topic, rather than deliberately preferential treatment for one particular point of view.

One of the few suggestions that a *Wikipedia* article showed bias of any kind came in one reviewer’s comments on the *Energia Renovable* article:

- “There is a tendency to disregard nuclear energy, particularly fusion, which is a flawed view commonly supported by green energy advocates.” (Reviewer 1 – Professor)

It was, in fact, more often the case that *Wikipedia* articles were credited with a distinct lack of bias, even with regard to topics where the risks of favouring one particular viewpoint (i.e. with respect to historical or political issues) might be considered to be quite high:

- “The second article [*Wikipedia*] was much more up to date, although not as much as it should have been. Both articles need to be updated – the first more so than the second. Given the ongoing political changes that are currently taking place in the Middle East, it is crucial to update these articles to reflect such pressing issues. Both articles were concise and fairly eloquent. The first one had more information, especially general information regarding climate and topography, etc. The second one focused almost entirely on political and economic issues in the Middle East. Nevertheless, the first one [*Mawsoah*] was much more biased and had a political tone to it, while the second one addressed such political issues from a seemingly objective point of view.” (Reviewer 1 – research student – *Middle East*)
- “a very controversial figure. I don’t really see how one could be much more neutral.” (Reviewer 1 – professor – *Hugo Chavez*)

5.2.3 Qualitative Judgments related to English, Spanish and Arabic Encyclopaedia Entries

As mentioned in Section 3, owing to the challenges of securing reviewers to participate in the study within the timescales, not all articles were evaluated by the same number of reviewers. Similarly, numbers of reviewers taking part in the study varied by language. There were eight reviewers for the Arab articles, eleven for the English articles and fourteen for the Spanish articles.

Additionally, owing to the difficulties of identifying a publication with a sufficiently wide spread of articles, two separate Arabic publications were used: *Mawsoah* and the *Arab Encyclopaedia*. Of those publications, the two articles taken from the *Arab Encyclopaedia* (on Algorithm and Mathematical proof) received very positive judgements. Each of these articles was evaluated by three reviewers, whereas two of the four articles compared to *Mawsoah* (which were generally less well received) were only evaluated by two reviewers.

Therefore, whilst the Arab reviewers tended to be more critical of *Wikipedia* articles than English or Spanish reviewers, such a count will not necessarily give an accurate picture.

Across subject domains, there are similarities in the criteria employed by reviewers across all three languages, for example in reviews of potentially controversial social science subjects which are subject to change over time, both Arabic and Spanish reviewers used wording relating to completeness, neutrality and currency in their assessment:

- “The first article [*Wikipedia*] was much more up to date and discussed issues that occurred in 2011, while the second did not discuss anything that occurred in the 2000s, which therefore made it not very useful [...] When it came to political issues, neither article was sufficiently critical.” (Reviewer 1 – research student – *Egypt*)

There was, though, some variation across the Arabic, Spanish and English examples with respect to opinions about the quality of language, where concern for traditional notions of language use appeared to be judged more critically by both Spanish and Arabic reviewers than was the case with English language articles:

- “Use of Spanish is alternating between Latin American Spanish and Spain Spanish.” (Reviewer 1 – professor – *Energia Renovable*)
- “The article is not well written; the organisation is poor, the verb tenses inconsistent and there are sections misplaced. Just to illustrate [...] what does a ‘relación sentimental e ideológica’ mean? Ideological is normally not used to refer to a relationship.” (Reviewer 3 – research student – *Hugo Chavez*)
- “[...] weak and poor Arabic language that can’t be understood.” (Reviewer 3 – student – *Pharmacokinetics*)

Overall, our analysis suggests that *Wikipedia* articles were judged favourably more often than not (in some cases just marginally, in others quite markedly) compared with articles from *Britannica*, *Enciclonet* and *Mawsoah*, but this was not the case when compared with *Arab Encyclopaedia*, whose articles were more often judged favourably than *Wikipedia* articles.

5.2.4 Qualitative Judgments Related to Different Disciplinary Areas

The disciplinary divisions selected for the study reflected those used to structure subjects at the University of Oxford, where the study took place. As a result, the project divided up academic disciplines according to the four main disciplinary ‘divisions’ of the University: Humanities; Social Sciences; Mathematics, Physics and Life Sciences; Medical Sciences. In reviewing the results of this study, we have found though that these disciplinary divisions, whilst helpful in logistical terms, did not constitute sufficiently clear disciplinary distinctions to be useful for analysis purposes. Therefore, we focused our attention in reviewing the qualitative data on disciplinary difference across two broad categories: 1) Humanities and Social Sciences, 2) Mathematics, Science and Medicine.

For the most part, reviewers in all subject areas worked well enough with the range of criteria against which they were asked to judge articles. In analysing their final comparative

comments about articles, it is possible perhaps to detect distinct tendencies to prefer slightly different values in the humanities and social science article reviews, from those of mathematics, science and medicine.

For instance, terms introduced by reviewers into their discussion of history and social science articles included 'polished', 'eloquent', 'aesthetic', 'scholarly' and 'coherent'. By contrast, the key notions valued by reviewers of mathematics and science articles seemed especially to be those of scientific thinking, clarity and – above all – conciseness:

- “The difference between the articles is very stark. The first is very waffly and never really gets on to the actual substance of what attention is / how it works. By contrast, the second article is concise and yet covers the important main points.”
(Reviewer 1 – academic – *Attention*)

Such fairly predictable and minor distinctions aside, though, it is not possible to add in any significant ways to the detailed analysis of the quantitative data concerning academic discipline variation as reported in 5.1 of this report.

6. Discussion

6.1 Methodology

Section 3 of this report describes in some detail the full process followed in carrying out this research. In particular it reported on the processes of selecting languages and encyclopaedias for comparison with *Wikipedia* articles, the sampling strategies for student and academic expert reviewers, the selection of articles and the review process.

We have nothing to add in terms of the decisions made for which languages to study, beyond saying that nothing which occurred subsequently in the process suggested that these were inappropriate choices. Considerable care was spent in trying to select the online encyclopaedias that were most appropriate for comparison with *Wikipedia*, in terms of the nature of content, style and readership. We have no doubt, in retrospect, that even given the difficulties in finding articles of equivalent focus and length to *Wikipedia* articles on a number of occasions, we could not have made better choices with respect to the three languages chosen.

The processes of establishing the samples of students and academic experts proved to be largely appropriate and productive. Achieving an initial pool of 116 students provided an excellent foundation for the selection of 24 students (12 as the main cohort, and a further 12 as back-up). Given the time pressures and commitments of such high level students, we were pleased that we were able to select this number of committed and capable people who had such a key role to play in the research, both in terms of identifying academic experts and in carrying out their own reviews of articles. The identification of experts was carried out rapidly and productively and resulted in a generally satisfactory outcome in terms of numbers and quality of reviews. However, we were not always able to meet our target of at least two academic experts for each article (as against one student and one academic expert, a minimal requirement that was met on every occasion). Given the considerable efforts and enthusiasm of all involved, especially the students, this does raise serious questions about the viability of a significantly large-scale study in the future.

Similar questions arise from the difficulties encountered in selecting and preparing pairs of articles for comparison. In the event, it proved extremely difficult to locate encyclopaedias which provided articles that could be compared with *Wikipedia* in a number of the specialist areas of experts. We had no alternative but to select topics that were broader and less specialist than we would have preferred and which did not match the expertise of academic reviewers as closely as originally intended. This, once again, has significant implications for any future scaling up of the research, although we do believe that the actual process of comparison proved to be extremely valuable (discussed further below).

Another potential problem in preparing articles concerns the inclusion of additional material such as photographs, charts and tables. Images, as was explained in Section 3, had been removed from articles and presented separately as part of the anonymisation process, so that although reviewers were not able to see images in context, they were able to comment on them. A few reviewers commented on the issue of imagery as a concern especially for

articles on science and social science. For instance, with respect to the *Wikipedia* article on *Neuronia*, Reviewer 1 commented approvingly about the presence of images (“It includes more photographs. These photographs help to understand the role of neuron”). This is a factor that helped to distinguish two articles judged to be generally of commensurate quality. In general, though, imagery was not frequently referenced in comparative comments, perhaps because images had been dislocated from the flow of the text.

It is worth noting that a general audience (rather than academic reviewers, who may well be more used to engaging with largely text-based material) might expect high quality imagery from an encyclopaedia, online or otherwise, and might make overall judgments of quality based on layout and the quality of the imagery included. The methodology of this study (removing images from their original location and using academic reviewers) meant that the focus was very strongly on the words used and may not have fully captured this area of judgment.

The greatest difficulty in anonymising articles involved the removal of information that may be considered integral to the value of the article, such as the *Wikipedia* article tree, or the name of the author of some articles in other encyclopaedias. The removal of such information was clearly essential in order to achieve the goal of blind reviewing, but it could be argued that information about authorship might, to some extent, compensate for lack of references (although there is no inherent reason why named authorship need preclude the use of references).

The review process appeared to have been productive and appropriate. The criteria contained within the feedback tool (whose development is described in Section 3) provided an appropriate range of distinct perspectives on articles and stimulated a range of judgments and comments that, for the most part, enabled us to gain quite a rich and insightful range of comments about articles from reviewers. However, in developing this tool further, we would recommend a further period of trialling of criteria, especially around concepts such as completeness, conciseness and coherence, which sometimes seemed to generate slightly contradictory comments from some reviewers.

There is, of course, a fundamental problem in trying to reconcile the provision of clear and consistent criteria so that a wide range of reviewers can be seen to be making comparable judgments, with the need (especially when it comes to asking for qualitative judgments) to capture the language and criteria that academic experts might otherwise have used, if simply asked to discuss the strengths and weaknesses of articles as they perceived them. It is certainly only through such an approach that it would be possible to carry out any systematic form of quantifiable content analysis of experts’ qualitative judgments. As it was, in analysing the qualitative aspect of reviewers’ judgments, analysts had to make their own judgments to some extent about whether, for instance, a reviewer talked about enjoyment of an article because that term had been put before them in the feedback tool or because they had actually enjoyed reading the article.

This said, we felt that the qualitative comments provided considerable insight into the kinds of criteria and standards for making judgments about online encyclopaedia content that different academics use, and it may have been the case that we would have had considerably more difficulty in generating the quality of thinking and comment that we did receive without the framework that was provided. It is fair to say, though, that this pilot has

not provided a definitive answer to this question, but it has demonstrated that the approach of providing a clear and specific framework can be highly productive. We encountered no evidence, at any rate, that reviewers felt constrained by the criteria provided for the review process. They were clearly capable of introducing their own criteria into the qualitative discussion of articles as appropriate, such as in the following:

- “The discussion of the Monologion is fairly good. There are some factual errors or at least infelicities: ‘monologium’ and ‘proslogium’ should be ‘monologion’ and ‘proslogion’. Anselm was most likely canonised in 1494. The treatments of the proslogion and cur deus homo are superficial.” (Reviewer 2 – *St Anselm*)

We definitely believe that the comparative approach worked very effectively, and we would certainly recommend using that more widely, all other considerations being equal. This was clearly demonstrated on a number of occasions by comments such as the following:

- “Reading the second article made me realise how poor the first article was in that it did not cover the subject comprehensively and focused excessively on one aspect that could be viewed as peripheral. In the second article, the structure and the reference to sources were ideal.” (Reviewer 1 – *antibiotic resistance*)
- “The most important differences separating the two articles were conciseness and scope of information.” (Reviewer 2 – *Mutation*)
- “Article 1 is superior in almost all respects to article 2. The difference is particularly striking in terms of structure, references, factual accuracy, grammar and language.” (Reviewer 2 – *Numero Racional*)

In addition to the way in which comparing articles managed to focus reviewers’ awareness of the qualities and weaknesses of specific articles, it was interesting to see on a number of occasions the way that comparison generated insights into the ways in which an article on a particular topic might usefully combine the insights and approach of multiple articles:

- “Although both articles address the same basic issue of global warming, each of them with very different views, I find both very instructive and entertaining. The first privileges the sociological items, while the second starts with a more geophysical outlook. I find them both very good and complementary.” (Reviewer 1 – *Cambio Climatico*)
- “I preferred the first article to the second. It is written in a more scholarly manner and it provides a lot of references. I found the second paper still a draft, and this might be the case. Ideally you would combine the two to give a more comprehensive picture of preschool education.” (Reviewer 2 – *Preschool Education*)

One methodological issue that raises perplexing questions is the fact that reviewers’ judgments were not always in exact agreement with one another on specific articles. To some extent, such differences reflect the previous point about the different perspectives on particular topics that emerge from different articles. We do not consider, certainly, that different viewpoints on the same topic are necessarily invalid – indeed, they are part and parcel of academic life, as is variation in academics’ judgment of quality. Just as an article submitted to a journal for peer review will very often receive diverging judgments, so did a number of the online encyclopaedia articles in the present sample (such as, for instance, the following where one reviewer differed markedly from the other reviewer or reviewers:

Energia Renovable, Antibiotic Resistance, Preschool Education, Egypt). For the most part, such disagreement concerned issues of emphasis and style rather than accuracy or perhaps in a small number of cases also reflected negative perspectives on online encyclopaedias in general. This raises questions that are not possible to resolve here, but which might need clarifying before further work of this kind is carried out. These concern philosophical and epistemological perspectives on issues such as: the nature of knowledge within different cultural settings, the traditional role of encyclopaedias as sources of authoritative knowledge, perspectives on the Internet in general (and *Wikipedia* in particular) as a medium for the co-construction and sharing of knowledge, and so on. It is a mistake, perhaps, to assume that everyone involved in a project of this kind is actually agreed on the fundamental perspectives, which are bound to influence judgments made about individual articles.

Finally, given the successful outcome in terms of return of reviews, we believe that the tool created for this pilot study, using Moodle, proved to be usable, and provides a good basis for further development. Indeed, we can say with some confidence that the decisions made for carrying out this pilot proved generally to be appropriate and effective, both in terms of securing the valuable co-operation of many busy academics within quite a tight timescale, and in terms of generating an illuminating and satisfactory dataset. We recognise that were the study to be substantially extended, some of the challenges in securing the necessary content for analysis and the desired range of reviewers might prove hard to surmount on a far larger scale.

6.2 Findings

The quantitative and qualitative findings from this project are more or less in agreement with one another, as might be expected. But they do lead to slightly differing perspectives on the judgments made by reviewers in one or two respects. While it is inevitable that quantitative results offer a more precise account of reviewers' judgments, we would suggest that the qualitative perspectives provided by the data are also of considerable value.

6.2.1 Quantitative Findings

The quantitative findings demonstrate that, across the piece, *Wikipedia* articles scored more highly on accuracy, amount and quality of references, style/ readability and overall judgment (which is to say, citability). With respect to citability, though, it must be emphasised that at no time did articles from online encyclopaedias, whether *Wikipedia* or others, score highly with respect to this key criterion. This was also made quite clear in the qualitative comments. While many reviewers felt that some of the online encyclopaedia articles they reviewed were suitable for use in non-academic contexts (as useful or interesting overviews and introductions on particular topics) they did not consider that such articles could be considered on equal terms to material in refereed journals or textbooks from established publishers. Indeed, for academic reviewers in general, this was not likely to be otherwise and should not be seen either as a particularly surprising outcome, or as a particularly negative reflection on such articles.

This simply reflects the reality that scholarly knowledge and scientific research have to go to far greater lengths than are possible within a relatively short encyclopaedia article in order

to justify knowledge claims in general. By ‘far greater lengths’, we should add, we are talking especially about issues such as extent of evidence provided in support of a knowledge claim, clarity about methodological issues and evidence of peer review. None of these things can reasonably be expected of articles in online encyclopaedias in sufficient measure. It is important here to focus on the qualities that can reasonably be expected of such sources of knowledge, in order to see whether – on the basis of this quite small sample at least – we were able to collect evidence which, if collected on a far larger scale, would provide definitive judgments about the quality of *Wikipedia* in its own terms, which is to say, as a leading online encyclopaedia.

Within this small sample, *Wikipedia* scored well in many key respects, as we have indicated above, and these positive scores were reflected when considering the findings in relation to the specific perspectives of articles in different languages, and in different disciplines. Indeed, as the quantitative results show clearly, it was only with respect to articles in the Arabic encyclopaedias that *Wikipedia* did not earn markedly higher scores. In the case of those two encyclopaedias, *Mawsoah* and *Arab Encyclopaedia*, *Wikipedia* came out lower on style, and more or less the same on the other key criteria of accuracy, references, overall judgment and overall quality score. In all other comparisons, *Wikipedia* fared somewhat better on references and, with the exception of articles in the Humanities and MPLS (Mathematics, Physics and Life Sciences) where *Wikipedia* scored no better on accuracy, style/ readability, overall judgment and overall quality score. This was more or less the case with articles in the Social Sciences, with the difference that *Wikipedia* scored relatively poorly there on style/ readability. In Medical Sciences, though, *Wikipedia* scored well on accuracy, references and overall judgment.

6.2.2 Qualitative Findings

In terms of qualitative analysis, the picture is less easy to summarise. It is, in theory, possible to total the number of positive and negative comments and the overall number of preferences expressed regarding the full spread of articles in the sample. However, reviewers were generally quite measured in their comments and sometimes expressed no distinct preference, or highlighted strengths and weaknesses across both articles whilst marginally preferring one. For some articles overall preference is too close to call and in others where a preference is expressed, it is not a strong preference. Additionally, some subjects had four reviewers, whereas others only had two, so any overall count of preferences will be necessarily skewed by this. If a particularly well received article from one publication happened to have more reviewers than a less well received article from the same publication, that publication would make an unrepresentatively strong showing in any rough total of reviewer preferences.

It would, at any rate, be pernicious to attempt to quantify qualitative judgments too precisely. Above all, the analysis of qualitative data aims to capture things that are hard to quantify precisely: feelings, attitudes and opinions of reviewers that are important and illuminating but are often also imprecise and hard to compare.

In comparing ‘the accuracy, quality, style, references and judgment of *Wikipedia* entries as rated by experts to analogous entries from popular online alternative encyclopaedias’ through the medium of the qualitative data, we were able to identify a number of issues

both about the way academics make judgments about online encyclopaedia content in general and in particular about the characteristics that distinguish *Wikipedia* entries. It is evident from these qualitative judgments that, apart from a small number of articles considered to be quite weak, *Wikipedia* articles in general emerge creditably from this comparison in a number of respects. More usefully, it was possible to identify a pattern of qualities that appeared to be particularly characteristic of *Wikipedia* entries. They were generally seen as being more up to date than others, were generally considered to be better referenced and appeared to be at least as strong as other sources in terms of comprehensiveness, lack of bias and even readability.

This latter judgment is worth emphasising here, because the quantitative data suggests that *Wikipedia* generally performed less well than the other sources when it came to style/readability. The qualitative data does show though, that despite the readability issues associated with multi-authorship, such as lack of cohesion, repetition and poor structure, there was no clear impression across the qualitative data that *Wikipedia* articles were all less satisfying or engaging to read than other articles. This should definitely be considered as a crucial aspect of readability in our opinion. In the reviews, for example, of the *Wikipedia* article on *Polinomia* (“it has the right encyclopaedic tone”), *Memory* (“concise and well-written”), *Attention* (“very carefully written”), it is clear that reviewers approved the reading experiences as much as they valued the accuracy and references. By the same token, many articles from other encyclopaedias were criticised for their poor quality of writing.

Nonetheless, the multi-authored nature of *Wikipedia* did frequently lead to negative judgments regarding repetition, poor structure and lack of coherence within articles. But the analysis of qualitative data did allow us to build a more interesting composite picture, in which the balance of qualities in a particular article were seen often to outweigh specific shortcomings. Insofar as *Wikipedia* articles were more often judged to provide more comprehensive and up to date content, useful references and at least comparable levels of accuracy and citability, it can be argued that reviewers (though critical when asked about readability) were prepared to forego that to some extent given the presence of the other qualities.

Indeed, in seeking to characterise what it was about a high proportion of the *Wikipedia* articles that led to them being preferred in the qualitative findings as a whole, we would suggest that the answer can be found in the marked impression that the strengths of multi-authorship were often judged to outweigh its weaknesses. Regardless of problems of readability, style and structure on occasions, the greater likelihood that it was a *Wikipedia* article that would be judged to be up to date, comprehensive and well referenced in the qualitative comments offers, at the very least, a hypothesis about the particular qualities of *Wikipedia* that is well worth exploring in a more substantive study.

6.3 Recommendations

6.3.1 Methodological Considerations for Future Research

We are well aware that the findings of this study are merely indicative of the kinds of approaches and issues that might be considered relevant to a future large-scale study, and we must repeat that although the findings for this small sample were quite positive with

regard to *Wikipedia*, this has limited significance beyond indicating the kinds of questions or hypotheses that might profitably be focused on in a subsequent study on a larger scale. Although it is perfectly legitimate to look in depth, as we have attempted to do, at the themes emerging from this small sample, in order to sharpen the focus of further research of this kind in the future, it is obviously not possible to make generalisations about *Wikipedia* as a whole on the basis of the comparative analysis of 22 articles. Therefore, it is important to consider what kinds of further study are most feasible.

The use of the feedback tool devised for this research produced results that were comparable across reviewers, and allowed for consistency of measurement of views, especially in terms of quantitative data. Therefore, in any future study, we consider that it would definitely be worthwhile to develop and refine this instrument. Further pilot work would be needed to test the appropriateness of specific criteria, such as conciseness, readability and enjoyment, in order that reviewers are consistent in their application of these to specific articles. Further consideration as to whether it would be appropriate to modify the tool for use in different disciplinary areas needs examination: our findings indicated that the notion of conciseness, for instance, was used differently by reviewers in science areas than in humanities and social science areas. It is also worth exploring whether an element of unstructured questioning of academics might produce more valuable qualitative data.

In terms of methods used, we do have considerable concerns about the feasibility of substantially scaling up the exact approach used in the present study. This is for a number of reasons:

1. The difficulty of finding a substantially increased number of articles from other encyclopaedias against which to compare *Wikipedia* articles.
2. The difficulty of securing and managing the participation of a sufficient number of academic experts for a large-scale representative study.
3. The complexity of carrying out a sufficiently rigorous analysis of qualitative data on a large-scale without recourse to some degree of quantitative content analysis, which is not feasible with the present model of feedback tool.

In other words, we would say that the present project has produced a considerable amount of interesting and illuminating data, but we recognise that there are serious logistical problems in replicating it on a far larger and wider scale. Although it would presumably be possible to do so, given considerable investment in research staff to carry this out across multiple sites, the question arises as to whether or not this would be worthwhile.

6.3.2 The Focus for Future Research

As the end of the Section 6.2 indicated, this study has indeed raised some lines of enquiry for further research of a similar kind.

As mentioned above, the qualitative results from this study indicated that academic reviewers are generally open-minded about what constitutes quality in articles in online encyclopaedias. We would therefore suggest that, on the basis of these findings, the following tentative hypothesis is worth testing in a future study: *the perceived quality of*

online encyclopaedia articles is as much dependent on a coherent and engaging narrative about a topic as it is on extensive provision of technical information.

Other issues that have arisen in this study that merit further study along the same lines also might include questions about the extent to which multi-authored articles present multiple, repetitive or even contradictory perspectives on their topics. Of importance also is the appropriateness of different kinds of reference to other sources, with respect to different kinds of content: is it possible to identify variation in terms of academic discipline/ topic in order to judge when it is appropriate to draw on internet-based references as well as, or instead of, references to peer-reviewed journals and published books?

We would also recommend considering a wider disciplinary focus to include academic topics that are particularly relevant to *Wikipedia's* distinctive strengths, such as disciplines around which there is substantial internet-based discourse and dissemination: cultural studies, information sciences, communications, journalism, media studies and a wide range of interdisciplinary studies that bring together areas such as: economics, geography, sociology, future studies and sociolinguistics.

In conclusion, though, we must again ask whether the considerable investment needed to carry out large-scale research into the question of academic approval for *Wikipedia* articles constitutes the best way forward. It is important that ongoing effort continues to be made to secure the judgments and engagement of academic experts in monitoring the quality of *Wikipedia*, but the experience of this study suggests that a subsequent study of this kind on a considerably larger scale is not necessarily the most appropriate way of achieving this.

Previous studies, such as those referred to in the Introduction to this report, have been inconclusive, and we would not claim that this study has demonstrated a feasible means of replicating or extending the findings of the original study by *Nature*, although we do believe that it has made a respectable addition to studies of that kind. The criteria and terms of reference of previous studies have never been consistent from one study to the next and we suggest that a possible way forward is to seek to establish a more consistent set of criteria and questions (drawing on the experience of the present study), that can form the basis for an continuing series of manageable, small-scale studies of quality in the future. These would perhaps form the basis of regular snapshots that help to monitor what is, inevitably, a continuously shifting picture.

While we recognise and applaud the efforts of *Wikipedia* to maintain high standards in all their articles, though, this study has indicated that even the highest standards are not likely to convince academics – quite reasonably, we suggest – that *Wikipedia* articles can expect to be citable alongside peer reviewed journal articles, or even published books. These academics were, on the other hand, very prepared and able to recognise the qualities and values in the best online encyclopaedia articles – the majority of which, in the present study, were found in *Wikipedia* – in their own right.

We suggest that it is worth considering also whether future research into *Wikipedia* might – in addition to the rigorous small-scale studies of accuracy suggested above – not usefully attempt to devote more attention to the ways in which this exceptional resource is being used by a wide range of readers – academics, students, workers in various industries and self-directed learners also – as a crucial source of knowledge. In this respect, questions of

considerable interest might include: What do they expect from it? What do they gain from it? What credit do they attribute to the accuracy of content they encounter? How do they follow up the openings to new knowledge that it provides? And so on. The academics who contributed to this research have very helpfully recognised that the quality of online encyclopaedias resides not so much in exhaustiveness of content, so much as in their capacity to make knowledge accessible and engaging to a wide readership.

Appendix I

- 1. Prototype for Selection of Spanish Encyclopaedias for Comparison with Spanish Wikipedia***
- 2. Student Biography and Nomination Sheet***
- 3. Article Feedback Questionnaire***
- 4. Article Comparison Questionnaire***
- 5. Declaration***
- 6. Example of Article Ready For review, After Completion of Standardisation and Anonymisation Process***

Appendix I (1): Prototype for Selection of Spanish Encyclopaedias for Comparison with Spanish Wikipedia

	Encyclopaedia Source	Existence of Article on the Topic	Article Length in Pages	Completeness of Article	Presence of Internal and/or External Links at the Bottom of Article
Politica de Bolivia	Wikipedia	Yes .	2	No – does not mention the history of Bolivian Politics, or the different political parties, or the different ministries (this is compared to the page on The Politics of Spain).	No – just has one external link.
	Enciclonet	No – there is no individual article about Bolivian politics, but it is mentioned on a general article about Bolivia.			No
		Evo Morales (Enciclonet).	5	There does seem to be a lot of information, though it is not well divided into sections, and the focus of the article is his life, and not the “politics of Bolivia”.	No
Hugo Chavez	Wikipedia	Yes	19	Yes – it mentions his life before becoming the Venezuelan President, and then his years as a President.	Yes – It has 2 “See also” links, which take you to Wikipedia pages, one of which is out of date.
	Enciclonet	Yes	7	Yes – it seems quite full of information, but I think to make readers follow the text easier it needs more subheadings.	No
Cambio Climatico	Wikipedia	Yes – takes you to a page titled “Calentamiento Global” (in the English Wikipedia there is a page for Global Warming, and one for Climate Change – because they are in fact two different things).	15	Yes – it has a lot of information the causes and the effects of global warming, and there is one section about temperature changes – rather than more broadly climate change.	Yes – it has 4 “See also links”.
	Enciclonet	Yes	4	Yes – it seems quite concise, shorter than the Wikipedia article, but a good introduction to climate change. It does not go into a lot of detail.	Yes – to related pages within Enciclonet.
Energia Renovable	Wikipedia	Yes	16	Yes – it does seem complete, plus it is good that each section of the different types of renewable energy sources has its own link to a Wikipedia article on it.	Yes – has 8 links in the “see also” section.

	Encyclopaedia Source	Existence of Article on the Topic	Article Length in Pages	Completeness of Article	Presence of Internal and/or External Links at the Bottom of Article
	Enciclonet	No – it has an article on “Energía Alternativa” .	1	Energía Alternativa. No - Very short, and conveys its ideas in a complicated way. It does not go into any detail.	Yes – to related pages within Enciclonet.
Numerical Analysis/ Análisis Numérico	Wikipedia	Yes	5	No – comparing it with the English version it needs more information regarding the history of Numerical Analysis, more on Generation and propagation of errors.	No – although it does have a section on “otros temas de análisis numérico” which includes links to different error types.
	Enciclonet	No page on it			
	Enciclonet	Yes – on Polinomio.	12	Yes- seems quite detailed and broken down into sections which makes it easier to follow.	Yes – one link to “algebra”.
Partial Differential Equation/ Ecuación en derivadas parciales	Wikipedia	Yes	4	No – compared with the English version The English version has more examples, the classification includes equations of first and second order (Spanish one only includes second order), and there is a section on analytical methods to solve PDEs and numerical methods to solve PDEs, which are not present in the Spanish article.	Yes – in the “see also” section, there are four links.
	Enciclonet	No – there is a section in the article “Ecuación diferencial” about “Ecuación diferencial en derivadas parciales”.	3	No – compared to the Wikipedia English page, but still seems to have some basic information.	No
	Enciclonet	Numero racional.	7	Yes – there is detail on it at a basic level.	No
Neural Network	Wikipedia	No – there is one on “Red neuronal artificial” so just on artificial neural network.	7	The article on Artificial neural networks does seem to be detailed, again not as much as the English version of it. However, there is no general article in Spanish on Neural Networks.	No
	Enciclonet	No – nothing of the sort.			
	Enciclonet	Neurona, Yes.	2	Not as much as there could be on a page about neurons. The information is very basic, descriptive, it does not go into details.	Yes – one link to the “nervous system”.
Proprioception	Wikipedia	Yes.	3	Yes – needs a section on applications, and a more detailed one on impairments of proprioception.	Yes – “see also” links.
	Enciclonet	Yes.	1	No – it is just a definition of the word, it does not go into any detail.	No

Appendix I (2) Student Biography and Nomination Sheet

University of Oxford

Wikipedia and Epic, UK



This form consists of 3 sections:

Section 1 consists of your biographical details

Section 2 consists of details of your current and past education/research, and information on your area of academic expertise within your discipline

Section 3 consists of nominating 4 academic experts/professionals working in your discipline and who will be able to review articles in your native language.

Further instructions are provided at the beginning of each section. Please follow these instructions carefully and complete all boxes. Please use a computer to fill in the boxes, handwritten forms will not be accepted. Please do not exceed the word limits specified for items.

PLEASE COMPLETE THIS FORM AND RETURN IT TO wikioxford.study@gmail.com BY 5:00 pm on 28 NOVEMBER 2011. FAILURE TO RETURN THIS FORM BY THIS DATE MAY RESULT IN EXCLUSION FROM FURTHER PARTICIPATION IN THE STUDY.

SECTION II: AREA OF ACADEMIC EXPERTISE

Area of Expertise (Please restrict this to 3 areas)	
Title of Thesis	
Focus of Doctoral Research Work (Please restrict this to 300 words)	
Previous training and expertise (Please restrict this to 200 words)	
Key Words that describe the topic of your doctoral research (Please provide 5 key words for your area of expertise)	

SECTION III: NOMINATION OF ACADEMIC EXPERTS/PROFESSIONALS

- Please nominate three academic professionals/experts who work in the sample discipline as you, and who are fluent in your **native language** (this does not necessarily have to be their native language, but they must be sufficiently fluent to be able to review articles in your native language).
- Please ensure that the academics you nominate **have worked/work closely** with you, and are well versed with your discipline and your area of expertise (they should preferably be someone in your current or past lab/department, a supervisor or a senior post-doctoral researcher).
- Please pay careful attention to this section and select the academics thoughtfully. It is important that you have a **good personal rapport** with them as we will ask you to contact them about this study.

Academic Expert I

Title	
Name	
Position	
Institution	
Email address	
Focus of research (Please provide a brief summary of less than 100 words)	
Area of expertise (Please provide 3 key words)	
Relationship to you	
Area of academic overlap between you and the nominated academic (Please provide 3 key words)	

Academic Expert II

Title	
Name	
Position	
Institution	
Email address	
Focus of research (Please provide a brief summary of less than 100 words)	
Area of expertise (Please provide 3 key words)	
Relationship to you	
Area of academic overlap between you and the nominated academic (Please provide 3 key words)	

Academic Expert III	
Title	
Name	
Position	
Institution	
Email address	
Focus of research (Please provide a brief summary of less than 100 words)	
Area of expertise (Please provide 3 key words)	
Relationship to you	
Area of academic overlap between you and the nominated academic (Please provide 3 key words)	

Appendix I (3) Article Feedback Questionnaire

Please complete the questionnaire.

Continue the form.

(*)Answers are required to starred questions.

Page 1 of 15 Strengths and weaknesses

1. Please give your initial impressions of the key strengths of this article.*



2. Please give your initial impressions of the key flaws of this article.*



(*)Answers are required to starred questions.

Page 2 of 15 Validity

3. Use the scale 1 – 5 to indicate the validity of information in this article.*

- 1 (low validity)
- 2
- 3
- 4
- 5 (high validity)

Please list any specific information that is invalid and comment on your judgement. For example, did you assess validity by your own experience or by reference to sources?*

(*)Answers are required to starred questions.

Page 3 of 15 Breadth and quality of references

Breadth of references

4. Use the scale 1 – 5 to indicate whether this article is sufficiently supported by references.*

- 1 (insufficient)
- 2
- 3
- 4
- 5 (sufficient)

If you selected 1 – 4 on the scale, then please comment on the references that are missing.
If you selected 5, then please justify your answer.*

Quality of references

5. Use the scale 1 – 5 to indicate the appropriateness of references included in this article.*

- 1 (inappropriate overall)
- 2
- 3
- 4
- 5 (appropriate overall)

If you selected 1 – 4 on the scale, then please comment on the references that are inappropriate. If you selected 5, then please justify your answer.*

(*)Answers are required to starred questions.

Page 4 of 15 Completeness

6. Use the scale 1 – 5 to indicate how much, if any, key information is omitted from this article.*

- 1 (a lot)
- 2
- 3
- 4
- 5 (none)

If you selected 1 – 4 on the scale, then please comment on what key information is omitted. If you selected 5, then please enter N/A.*

(*)Answers are required to starred questions.

Page 5 of 15 Conciseness

7. Use the scale 1 – 5 to indicate whether this article is concise (both in terms of length and a lack of repetition).*

- 1 (lacks conciseness)
- 2
- 3
- 4
- 5 (concise)

If you selected 1 – 4 on the scale, then please comment on any particular extracts from the article that lack conciseness. If you selected 5, then please justify your answer.*

(*)Answers are required to starred questions.

Page 6 of 15 Relevance

8. Use the scale 1 – 5 to indicate whether overall the information in this article is relevant to the topic.*

- 1 (lacks relevance)
- 2
- 3
- 4
- 5 (highly relevant)

If you selected 1 – 4 on the scale, then please comment on any particular extracts from the article that lack relevance. If you selected 5, then please justify your answer.

(*)Answers are required to starred questions.

Page 7 of 15 Neutrality

9. Use the scale 1 – 5 to indicate whether this article is objective and unbiased. For example, you may consider if all salient aspects of the topic have been given appropriate importance, and if any controversies or gaps in knowledge are referred to.*

- 1 (subjective and biased)
- 2
- 3
- 4
- 5 (objective and un-biased)

If you selected 1 – 4 on the scale, then please comment on any particular extracts from the article that are subjective and/or biased. If you selected 5, then please justify your answer.

(*)Answers are required to starred questions.

Page 8 of 15 Currency

10. Use the scale 1 – 5 to indicate how much, if any, information is out of date in this article.*

- 1 (all)
- 2
- 3
- 4
- 5 (none)

If you selected 1 – 4 on the scale, then please comment on any particular information in the article that is out of date. If you selected 5, then please enter N/A.*

(*)Answers are required to starred questions.

Page 9 of 15 Well-written

11. Use the scale 1 – 5 to indicate whether this article uses clear and appropriate language.*

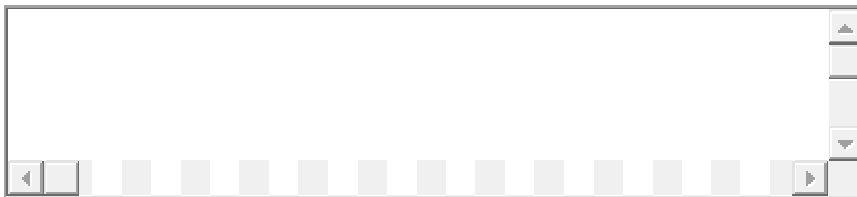
- 1 (unclear and inappropriate)
- 2
- 3
- 4
- 5 (clear and appropriate)

If you selected 1 – 4 on the scale, then please comment on any particular extracts from the article that are written without clarity, and/or inappropriately. If you selected 5, then please justify your answer.*

12. Use the scale 1 – 5 to indicate the level of accuracy of spelling, grammar and punctuation in this article.*

- 1 (poor)
- 2
- 3
- 4
- 5 (excellent)

If you selected 1 – 4 on the scale, then please give particular instances from the article where there are spelling, grammar and/or punctuation errors. If you selected 5, then please enter N/A.*



(*)Answers are required to starred questions.

Page 10 of 15 Clarity and organisation

13. Use the scale 1 – 5 to indicate how well-structured this article is, in terms of the order of information.*

- 1 (poorly structured)
- 2
- 3
- 4
- 5 (excellently structured)

If you selected 1 – 4 on the scale, then please comment on any particular extracts from the article that are not well-structured in terms of order of information. If you selected 5, then please justify your answer.*



14. Use the scale 1 – 5 to indicate how well-structured this article is, in terms of coherence between different paragraphs and sections.*

- 1 (poorly structured)
- 2

- 3
- 4
- 5 (excellently structured)

If you selected 1 – 4 on the scale, then please comment on any particular extracts from the article that are not well-structured in terms of coherence. If you selected 5, then please justify your answer.

(*)Answers are required to starred questions.

Page 11 of 15 Inclusion and integration of photographs and diagrams

16. Use the scale 1 – 5 to indicate how much, if at all, photographs and diagrams contribute to an understanding of the topic of this article. (Note: if there are no photographs or diagrams, then please select 'not applicable'.)*

- 1 (not at all)
- 2
- 3
- 4
- 5 (a lot)
- not applicable

Please comment on how photographs and diagrams do or do not contribute to an understanding of the topic. If there are no photographs or diagrams, then please comment on if, in your view, there should have been, or enter N/A.*

(*)Answers are required to starred questions.

Page 12 of 15 Enjoyment

16. Use the scale 1 – 5 to indicate how much, if at all, you gained enjoyment from reading this article.*

- 1 (not at all)
- 2

- 3
- 4
- 5 (lots)

Please comment on what you particularly enjoyed or did not enjoy of this article.*



(*)Answers are required to starred questions.

Page 13 of 15 Citability

17. Would you cite this article in a non-academic piece of work (e.g. for a college magazine or group newsletter)?*

- Yes
- No

Please justify your answer.*



18. Would you cite this article in an academic piece of work (e.g. for a peer-reviewed journal)?*

- Yes
- No

Please justify your answer.*



(*)Answers are required to starred questions.

Page 14 of 15 Improvements

19. Please comment on any ways this article might be improved that are not mentioned in your answers to previous questions.*

An empty rectangular text input box with a thin black border. It features a vertical scrollbar on the right side and a horizontal scrollbar at the bottom, both with a light gray and white checkered pattern.

(*)Answers are required to starred questions.

Page 15 of 15 Overall feedback and additional comments

20. Please use the space below to provide any overall feedback and further comments on this article.*

An empty rectangular text input box with a thin black border. It features a vertical scrollbar on the right side and a horizontal scrollbar at the bottom, both with a light gray and white checkered pattern.

Thank you for completing feedback for this article. Please move on to the next questionnaire.

Appendix I (4) Article Comparison Questionnaire

Comparative Question

Please answer this question.

[Click here to begin ...](#)

(*)Answers are required to starred questions.

(one) Please use the space below to make any additional comments about the two articles in comparison with each other.*

An empty rectangular text input box with a thin black border. It features a vertical scrollbar on the right side and a horizontal scrollbar at the bottom, both with a light gray and white checkered pattern.

Appendix I (5) Declaration

Declaration

Please complete the declaration form.

[Click here to begin ...](#)

(*Answers are required to starred questions.

1

(one) Please confirm that you have completed the Feedback Questionnaire for each article.*

Yes

2

(two) Please confirm that all reviews were conducted by you.*

Yes

3

(three) Please confirm that you did NOT make any attempts to identify the sources of the articles.*

No, I did not.

Appendix I (6) Example of Article Ready For Review, After Completion of Standardisation and Anonymisation Process.

Percepción

La **percepción** es un proceso nervioso superior que permite al organismo, a través de los sentidos, recibir, elaborar e interpretar la información proveniente de su entorno y de uno mismo.

Historia

Los primeros estudios científicos sobre percepción no comenzaron sino hasta el siglo XIX. Con el desarrollo de la fisiología, se produjeron el primer modelo que relacionaban la magnitud de un estímulo físico con la magnitud del evento percibido, a partir de lo cual vio su surgimiento la psicofísica.

Los personajes más relevantes en el estudio de percepción fueron:

- Hermann von Helmholtz, médico y físico alemán que realizó experimentos de acústica y oftalmología, entre muchas otras cosas.
- Gustav Theodor Fechner, psicólogo alemán autor de la ecuación que explica la relación entre el estímulo físico y la sensación (la llamada ley de Weber-Fechner]])
- Ernst Heinrich Weber, psicólogo y anatomista alemán fundador de la psicofísica.
- Wilhelm Wundt, médico alemán fundador del primer laboratorio de psicología experimental.
- Stanley Smith Stevens, psicólogo estadounidense autor de la llamada función potencial de Stevens.
- Max Wertheimer, Kurt Koffka y Wolfgang Köhler, psicólogos alemanes fundadores de la teoría de la Gestalt.
- Irving Rock, científico cognitivo estadounidense.
- David Marr, neurocientífico británico especialista en procesamiento visual.
- James J. Gibson, psicólogo estadounidense especialista en percepción visual.

Áreas

Los principales campos investigados en percepción se asemejan a los sentidos clásicos, aunque esta no es una división que se sostenga hoy en día: visión, audición, tacto, olfato y gusto. A estos habría que añadir otros como la propiocepción o el sentido del equilibrio. Tipos:

- Percepción visual, de los dos planos de la realidad externa, (forma, color, movimiento)
- Percepción Espacial, de las tres dimensiones de la realidad externa,(profundidad)
- Percepción Olfativa, de los olores,
- Percepción Auditiva, de los ruidos y sonidos,
- Cenestesia, de los órganos internos,
- Percepción Táctil, que combina los sentidos de la piel (presión,vibración, estiramiento)
- Percepción térmica, de las variaciones de temperatura (calor, frío)
- Percepción del dolor, de los estímulos nocivos,
- Percepción Gustativa, de los sabores,
- Quimioestesia, de los sabores fuertes, no se encuentra comprometida en caso de lesión de las áreas gustativas u olfativas
- Percepción del equilibrio
- Kinestesia, de los movimientos de los músculos y tendones
- Percepción del Tiempo, del cambio.Percibir implica la existencia de una reacción a una estimulación presente. Esta reacción se puede analizar en planos fisiológico, de consciencia o de conducta.
- Percepción de la Forma

- Percepción del campo magnético

Naturaleza de la percepción

La percepción es el primer proceso cognoscitivo, a través del cual los sujetos captan información del entorno, la razón de ésta información es que usa la que está implícita en las energías que llegan a los sistemas sensoriales y que permiten al individuo animal (incluyendo al hombre) formar una representación de la realidad de su entorno. La luz, por ejemplo codifica la información sobre la distribución de la materia-energía en el espacio-tiempo, permitiendo una representación de los objetos en el espacio, su movimiento y la emisión de energía luminosa.

A su vez, el sonido codifica la actividad mecánica en el entorno a través de las vibraciones de las moléculas de aire que transmiten las que acontecen en las superficies de los objetos al moverse, chocar, rozar, quebrarse, etc. En este caso son muy útiles las vibraciones generadas en los sistemas de vocalización de los organismos, que transmiten señales de un organismo a otro de la misma especie, útiles para la supervivencia y la actividad colectiva de las especies sociales. El caso extremo es el lenguaje en el hombre.

El olfato y el gusto informan de la naturaleza química de los objetos, pudiendo estos ser otras plantas y animales de interés como potenciales presas (alimento), depredadores o parejas. El olfato capta las partículas que se desprenden y disuelven en el aire, captando información a distancia, mientras que el gusto requiere que las sustancias entren a la boca, se disuelvan en la saliva y entren en contacto con la lengua. Sin embargo, ambos trabajan en sincronía. La percepción del sabor de los alimentos tiene más de olfativo que gustativo. Existe en realidad como fenómeno psíquico complejo, la percepción, el resultado de la interpretación de esas impresiones sensibles por medio de una serie de estructuras psíquicas que no proceden ya de la estimulación del medio, sino que pertenecen al sujeto. En la percepción se encuentran inseparablemente las sensaciones con los elementos interpretativos.

El llamado sentido del tacto es un sistema complejo de captación de información del contacto con los objetos por parte de la piel, pero es más intrincado de lo que se suponía, por lo que Gibson propuso denominarle sistema háptico, ya que involucra las tradicionales sensaciones táctiles de presión, temperatura y dolor, todo esto mediante diversos corpúsculos receptores insertos en la piel, pero además las sensaciones de las articulaciones de los huesos, los tendones y los músculos, que proporcionan información acerca de la naturaleza mecánica, ubicación y forma de los objetos con los que se entra en contacto. El sistema Háptico trabaja en estrecha coordinación con la quinestesia que permite captar el movimiento de la cabeza en el espacio (rotaciones y desplazamientos) y combinando con la propiocepción, que son las sensaciones antes mencionadas, relacionadas con los músculos, los tendones y las articulaciones, permite captar el movimiento del resto del cuerpo, con lo que se tiene una percepción global del movimiento corporal y su relación con el contacto con los objetos.

El proceso de la percepción, tal como propuso Hermann von Helmholtz, es de carácter inferencial y constructivo, generando una representación interna de lo que sucede en el exterior al modo de hipótesis. Para ello se usa la información que llega a los receptores y se va analizando paulatinamente, así como información que viene de la memoria tanto empírica como genética y que ayuda a la interpretación y a la formación de la representación.

Este es un modelo virtual de la realidad que utiliza la información almacenada en las energías, procedimientos internos para decodificarlas e información procedente de la memoria que ayuda a terminar y completar la decodificación e interpreta el significado de lo recuperado, dándole significado, sentido y valor. Esto permite la generación del modelo.

Mediante la percepción, la información recopilada por todos los sentidos se procesa, y se forma la idea de un sólo objeto. Es posible sentir distintas cualidades de un mismo objeto, y mediante la percepción, unir las, determinar de qué objeto provienen, y determinar a su vez que este es un único objeto.

Por ejemplo podemos ver una cacerola en la estufa. Percibimos el objeto, su ubicación y su relación con otros objetos. La reconocemos como lo que es y evaluamos su utilidad, su belleza y su grado de seguridad. Podemos oír el tintineo de la tapa al ser levantada de forma rítmica por el vapor que se forma al entrar en ebullición el contenido. Olemos el guiso que se está cocinando y lo reconocemos. Si la tocamos con la mano percibimos el dolor de la quemadura (cosa que genera un reflejo que nos hace retirar la mano), pero también el calor y la dureza del cacharro. Sabemos donde estamos respecto al objeto y la relación que guarda cada parte de él respecto a ella. En pocas palabras, estamos conscientes de la situación.

Entonces, como se indicó antes, la percepción recupera los objetos, situaciones y procesos a partir de la información aportada por las energías (estímulos) que inciden sobre los sentidos.

Para hacer más claro esto veamos el caso de la visión. Este sistema responde a la luz, la reflejada por la superficie de los objetos. Las lentes del ojo hacen que, de cada punto de las superficies visibles, esta se vuelva a concentrar en un punto de la retina. De esta forma cada receptor visual recibe información de cada punto de la superficie de los objetos. Esto forma una imagen, lo cual implica que este proceso está organizado espacialmente, pues la imagen es una proyección bidimensional del mundo tridimensional. Sin embargo, cada receptor está respondiendo individualmente, sin relación con los demás. Esa relación se va a recuperar más adelante, determinando los contornos y las superficies en su configuración tridimensional, se asignarán colores y textura y percibiremos contornos no visibles. Se estructurarán objetos y estos serán organizados en relación unos con otros. Los objetos serán reconocidos e identificados.

Este proceso se dará con la constante interacción entre lo que entra de los receptores, las reglas innatas en el sistema nervioso para interpretarlo y los contenidos en la memoria que permiten relacionar, reconocer, hacer sentido y generar una cognición del objeto y sus circunstancias. Es decir se genera el modelo más probable, con todas sus implicaciones para el perceptor.

La percepción está en la base de la adaptación animal, que es heterótrofa. Para poder comer las plantas u otros animales de los que se nutren, los animales requieren de información del entorno que guíen las contracciones musculares que generen la conducta, que les permite acercarse y devorar a su presa (planta o animal).

De este modo, la simple respuesta a las sensaciones, es decir al efecto directo de los estímulos, no fue suficiente; la evolución desarrolló paulatinamente formas de recuperar la implicación que tenían los estímulos en relación a los objetos o procesos de los que provenían; formándose así los procesos perceptuales.

Al contar con un sistema nervioso eficiente, este se empieza a usar para otras funciones, como el sexo, la sociabilidad, etc. Por ello, la percepción es un proceso adaptativo y base de la cognición y la conducta.

Bibliografía

- Merleau-Ponty, M. (1985). *Fenomenología de la percepción*. Barcelona: Planeta-Agostini. ISBN 84-395-0029-555.
- Bruce Goldstein, E. (2006 (2002)). *la percepcion del movimiento*(6º edición). Thomson. ISBN 84-9732-388-2.