NEW DIRECTIONS IN THE QUALITY CONTROL OF EPIDEMIOLOGICAL LECTURES
ON THE INTERNET

by

Faina Linkov

BS, University of Pittsburgh, 1999

MPH, University of Pittsburgh, 2001

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Faina Linkov

It was defended on

April 7, 2005

and approved by

Dissertation Director
Ronald E. LaPorte, PhD
Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Deborah Aaron, PhD
Assistant Professor
Department of Epidemiology
University of Pittsburgh

Sati Mazumdar, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Thomas J. Songer, PhD
Assistant Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Francois Sauer, MD
Supercourse Consultant
4513 W. 140th Street
Leawood, KS 66224-3632

Ronald E. LaPorte

NEW DIRECTIONS IN THE QUALITY CONTROL OF EPIDEMIOLOGICAL LECTURES
ON THE INTERNET

Faina Linkov, PhD

University of Pittsburgh, 2005

Finding high quality materials for the preparation of epidemiological lectures is a serious challenge for epidemiologists and public health professionals across the world. The emergence of the Internet in the early 90's offered a way to ease the access to the epidemiological lectures; however it also raised important questions about the quality of the educational lectures which are freely available on the Internet. In this research, we analyzed the quality of epidemiological lectures in the Global Health Network Supercourse lecture library.

We selected a random sample of 100 lectures in the Supercourse that accumulated at least 3 reviews from the visitors of the Supercourse sites. We found 7 experts, leading researchers in the field of public health and medicine, who were also very experienced in reviewing papers for journals. These experts evaluated the same set of 100 lectures and gave us their expert opinion on their quality.

Overall, the lectures were rated positively by both expert and the Supercourse reviewers. Although t-test indicated that the difference between the means was statistically significant, this difference is not meaningful due to large sample size. Kappa statistic and intraclass correlations indicated that inter rater agreement for experts and non-experts was surprisingly low (less than 0.4). We also observed HALO affect with overall score being a good predictor of other scores.

Our findings were consistent with existing research in the area of peer review, demonstrating low inter rater agreement. This poor inter rater agreement was demonstrated for

the first time for the Internet lectures. Our findings suggested that questionnaires assessing the quality of the Internet lectures may actually be replaced by one rating, similar to the system utilized in Amazon.com or hotel ratings.

This research was significant for the field of public health because it was one of the first efforts to evaluate the quality of epidemiological lectures on the Internet. The quality of lectures on the web has rarely been assessed scientifically for epidemiological and public health lectures. Future research in this area may need to concentrate on alternatives to the peer review system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

1. INTRODUCTION

"Quality is never an accident; it is always the result of intelligent effort."

John Ruskin (1819 - 1900)

Finding high quality materials for the preparation of epidemiological lectures is a serious challenge for epidemiologists and public health professionals across the world. Assistant professors teaching introductory epidemiology courses are forced to recreate simple "Epidemiology 101" lectures from scratch every time they start teaching a new course. Creating a brand new lecture every time results in a loss of time on the part of the educator, and often the resulting lecture is of poor quality, as it is built from new materials instead of existing strong materials. Typically, it takes fifteen to twenty hours to prepare a new lecture. This is a significant disadvantage for both new professors and students, considering how many high quality lectures have already been created by experienced researchers and instructors in the area of epidemiology, but are not reused. The process of new lecture preparation can be compared to the process of reinventing the wheel. Teaching experiences for new instructors often follow this unfortunate scenario, where they start out with no lectures to work with. Education might be markedly enhanced if there were mechanisms for obtaining template high quality, low cost epidemiological lectures available in one place, a lecture library for use by other faculty.

The emergence of the Internet technologies in early 90's offered a means to ease the access to epidemiological lectures. The number of Internet users and Internet sites has grown in a geometrical progression, with over 605 million people browsing the Internet at the end of 2002 (Nua Internet Surveys 2002). Epidemiology is increasingly present on the Internet, with over

100,000 epidemiological websites and over 100 websites of peer reviewed journals in the area of epidemiology.

Epidemiological research information is rapidly communicated through hundreds of epidemiological chat rooms, newsgroups, listservers, newsletters e-mails, etc. An epidemiology research education interface is rapidly developing on the web with more and more research data becoming available online each day. A considerable number of epidemiological materials are available online; yet there are no standards for quality. The Internet currently has over five million files in PowerPoint format, with over twenty thousand of them in the area of epidemiology (based on a search using the Google search engine). The number of epidemiological lectures on the Internet doubled in the past year and this number continues to grow. These lectures could be highly valuable to professors, teachers, and doctors worldwide, if there was a way to judge their quality.

It is not difficult to establish a lecture library on the Internet. One of the main difficulties associated with the development of a lecture library is the mechanism of quality control, as some of the lectures on the Internet may have inaccurate content or be outdated. The problem is that, in spite of the importance to our knowledge, there is no literature on quality control that we could find, despite the burgeoning growth of lectures on the web, and of Internet lecture libraries. We need to examine research literature related to quality control in peripheral areas that are related to internet materials and educational evaluation. It needs to be pointed out this literature is not closely related to the topic of the dissertation as there is no literature directly germane to this topic.

The most comprehensive and perhaps the simplest definition of quality is that used by advocates of total quality management (W. Edwards Deming 1982): "Doing the right thing right,

right away." When the expression "quality" is used, we usually think in terms of an excellent product or service that fulfills or exceeds our expectations. Quality control is the use of techniques and activities to achieve, sustain, and improve quality of a product or service (Besterfield 2001). One of the latest definitions of quality control (Barkman 1989) defines quality as "a measure of goodness that relates to the intended use of a product and the expectations customers have concerning this product".

The history of quality control is as old as the industry itself. The concept of labor specialization that was introduced during the industrial revolution resulted in the development of quality control discipline (Dhillon 1985). In 1950, W. Edwards Deming gave a series of lectures on statistical methods of quality control to Japanese engineers. Using these methods the Japanese set the quality standards for the rest of the world to follow. A quality renaissance began to occur in U.S. products and services in the late 1970's and 1980's, when the concepts of Total Quality Management (TQM) were publicized (Besterfield 2001).

Quality control is ubiquitous in industry under such terms as Statistical Quality Control and more recently Six Sigma (Westgard 2001). Statistical Quality Control, the branch of TQM, is the collection, analysis, and interpretation of data for use in quality control activities (Besterfield 2001). Six Sigma, utilized as a measure of quality at thousands of organizations around the globe, simply means a measure of quality that strives for near perfection. Six Sigma is a disciplined, data-driven approach and methodology for eliminating defects (driving towards six standard deviations between the mean and the nearest specification limit) in any process -- from manufacturing to transactional and from product to service (www.isixsigma.com).

There have been numerous scientific studies in industry evaluating quality control. According to Hilsenbeck et al. 1985, the principles of quality control exist to set the standards, maximize

reliability, reduce the sources of error, etc. A peer review system in the area of scientific publications and grant proposals also represents a form of quality control, however it is difficult to use for scientific lectures, due to its high cost, low throughput[1]*, and lack of information as to if they are validly measuring quality. Despite the fact that millions of articles and grants are subjected to QC each year, there is little data scientifically evaluating the process of traditional peer review in the framework of the science of Quality Control. Quality control is needed for Internet materials, especially for epidemiological lectures, but there are no accepted and tested means. The proposed research was one of the first efforts in this area.

Previous research studies have raised red flags about the quality of the biomedical information and epidemiological information on the web, because both misleading and life threatening advice is readily available from untrustworthy Internet sites (Impiccatore et al. 1997, Weisbord et al. 1997). A figure of 1400 "suspicious" websites was reported by one of the research studies in 1999 with a 21% increase in that number annually (Rogers 1999), and a recent US study found errors and contradictions even within sites (Berland 2001). The existence of these questionable epidemiological materials demonstrates that the development of new quality control mechanisms on the Internet is very important.

In this effort, we are not trying to create a traditional peer review system for Internet based epidemiological lectures, as we are not targeting journals. Various problems have been associated with the traditional peer review processes in biomedical journals. Although the Vancouver Group of Editors defines a Peer-Reviewed Journal (International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. http://www.icmje.org/) as: "**A peer-reviewed journal is one that has submitted most**

---

[1] In computer technology, throughput is the amount of work that a computer can do in a given time period. Throughput can also be defined as the speed of data transmission on the Internet. Historically, throughput has been a measure of the comparative effectiveness of large commercial computers that run many programs concurrently.

of its published articles for review by experts who are not part of the editorial staff", less than 50 % of papers in the leading biomedical journals like Lancet and Nature are peer reviewed. Our review of The Lancet revealed that less than 25% of the articles are peer reviewed. There appears to be a major disconnect where scientists see peer reviewed journals as being peer reviewed from cover to cover, but editors consider a referred journal as having only half the pages peer reviewed. A large body of papers, including correspondence letters, invited reviews, and editorials, published in leading journals such as Lancet are not reviewed externally The same is true for most of the upper tier scientific journals where less than 50% of the articles are peer reviewed.

Additionally, the science behind peer review mechanisms has not been explored. Editorial peer review, although widely used, is largely untested and its effects are uncertain (Jefferson et al 2002). Traditional peer review is not expected to work well for Internet lectures due to high cost and low throughput. Also, despite the 200 year history of peer review there has been virtually no scientific evaluation of the whole process. Thus, even today, we do not know if it works. The system of quality control we are setting up in this project is aiming to provide reviews for nearly all lectures at minimal cost.

The Supercourse is a library of over 2156 epidemiological and public health lectures (as of March 10, 2005), targeting the educator. It is a project based in the University of Pittsburgh, Department of Epidemiology and supported by the National Library of Medicine of the NIH. 20,300 researchers of the Global Health Network Supercourse project, from over 150 countries of the world are working together to share their best lectures in the area of epidemiology and public health in the Supercourse. It is one of the first efforts in epidemiology and health to target educators with web PowerPoint lectures, instead of students or consumers.

Although there are several studies looking at the quality of consumer oriented medical information on the web, none evaluate approaches for quality control of lectures. Despite its obvious importance, there are to our knowledge no research studies looking at the quality of the materials targeting the educator on the web. As described earlier, we could not find materials directly describing quality control of web based lectures. Several related areas were reviewed: web based quality control, web based peer review, automatic quality control, and educational evaluation, as they provide guidance as to what could be included in the investigation of quality control of epidemiological lectures on the Internet. The quality of lectures on the web has rarely been assessed scientifically for epidemiological and public health lectures. This project is one of the first efforts to evaluate the quality of epidemiological lectures on the Internet. In this project, we analyzed the results of web based reviews of the Supercourse lectures in comparison to reviews provided by experts and explored the applications of these findings to other fields. We hypothesized that overall positive ratings of the Supercourse lectures would be similar to the ratings of epidemiology experts, thus validating the process of Internet based peer review.

In the literature review section, we established the importance of quality control for epidemiological materials on the web, discussed existing approaches to quality control on the web and in the field of education in general, and suggested questions for the proposed research.

2.   BACKGROUND

"Focus on the Future"

(One of the concepts of total quality management)

2.1.   Internet and Biomedical Science

This past decade introduced the public to the concept of the Internet which gave professors, doctors, and the general public the opportunity to exchange information much more efficiently than ever before. The Internet revolutionized the way we thought about information. Suddenly, it became possible to rapidly share large volumes of information between continents and with minimal expenses, e.g. the death of distance. Professors from various universities, including medical and public health schools, have improved access to a variety of the latest research developments, via the Internet. Such valuable sources of information include search engines, free electronic journals, open source lectures, electronic books etc. Over the last several years, PubMed Central, BioMed Central, and the Public Library of Science have joined the older PubMed in providing much better access to scientific literature.

The importance of health information sharing and the Internet is something that cannot be ignored neither for faculty members around the world, nor for the consumers of health information. According to the latest estimates, approximately one third of consumers in England, and half of consumers in the U.S. rely on the Internet as a source of medical information (Eaton 2002). It is likely that the shift is occurring even more dramatically with scientists, as many of them now "Google" first and perform medline search second, if at all. This is to become even more frequent now with the advent of the Google Scholar.

7

2.2. Advantages and disadvantages of using the Internet for information exchange and education in the area of health: Educators as filters of information

Ever since health information became available on the Internet, various publications have addressed the advantages and limitations of using the Internet for obtaining health information. The Internet offers a unique and cost effective means to bring information across the digital or information divide. For places where biomedical journals are difficult to obtain due to high cost, it is now possible to obtain recent health information through the Internet. Several open access journals, including the Journal of Medical Internet Research, are now available free of charge on the Internet. Free previews of dissertations are now available through the digital dissertation database. MIT is making its course content freely available, etc. However, since anyone can set up a web site, there is a risk that, through ignorance or bias, the content of the site may not be correct even if the original information sources were reliable (Wyatt 1997). We reviewed a large number of papers describing the advantages and the limitations related to the use of the web based materials. We need to point out that these papers targeted websites highlighting materials targeting consumers, not educators. If incorrect information gets to consumers, there is likely a higher potential of adverse outcomes than for more informed educators.

Our approach is quality control of the lectures targeting the educators. It is our belief that targeting the educators adds another level of screening and quality control, and may be the most important quality control system. Educators are experts in their areas and thus are better adapt to tell a difference between materials of high quality and materials of low quality. Educators can serve as filters and prevent poor quality information from reaching consumers. If, after undergoing quality control procedures, a low quality lecture reaches an educator, the educator may easily disregard this information and not include it in his or her course as they know the area, and are thus are "very informed consumers." For example, an expert on diabetes

epidemiology is able to differentiate between good and bad materials in his or her area. Even when an expert is confronted with materials of poor quality, parts of such materials may still be invaluable (especially when it comes to research materials from the developing world). When materials go through educators, students are much less likely to learn from inadequate materials. Even though quality control of web materials is one of the few areas similar to the quality control of lectures, there are still major differences, including that the target audience has different educational needs and therefore different requirements for quality control.

Limitations of health related Internet websites in the fields of clinical medicine include poor quality information that may be dangerous for someone who is seeking health advice. Impicciatore et al. 1997 showed that parents searching for information about treating a feverish child could either receive good advice or be advised to administer aspirin, putting their child at risk of Reye's syndrome, according to which web site they visited. Safety of health information on the web has been compared to the safety of drugs 40 years ago when drugs were unregulated with regard to safety and efficacy (Rigby et al. 2001).

One of the obvious dangers of health informatics services cited by the literature is miscalculations of risks and false negatives/false positives given by internet sites or software. One of such examples was a miscalculation of Down's Syndrome risk for pregnant women (Cavalli 1996, Wilkinson 2000). One of the studies that looked at the accuracy of websites for managing fever in young children found that only a few web sites provided complete and accurate information and warned about potential risks of using the Internet to obtain medical advice (Impiccatore 1997). In both of the above cases, authors of the articles served as "educators", filtering out materials of poor quality. In both of these cases, the problem would not have happened if educators had previewed the information prior to publication. Despite some of

the potential drawbacks associated with the use of Internet materials on health, many patients, especially those fighting cancer, find the Internet to be an invaluable resource for health information in the privacy of their own homes (Ziebland 2004).

In 2001, Rigby concluded that the risk in health informatics services depends on a combination of the type of user, circumstances of use, type of use, and the nature of the system. According to Rigby, experienced clinicians may filter out spurious results received though incorrectly functioning computer based diagnostic support tools. In this example, targeting the correct type of user markedly reduces the risk of using computer technologies and adds additional level of quality control.

In the mid 1990's suggestions were made about giving accreditation or a stamp of approval to the websites that meet certain criteria (Forsström 1997), however this idea turned out to be impractical as the amount of websites and web based slides sprouted. Also, it was unclear as to who would give the certificate. Low cost automated quality control and quality control on the level of the user is becoming a topic of current Internet research. The work carried out in the Supercourse project suggested that targeting the educator instead of the consumer may be one of the ways to add another level of filtering / quality control to the evaluation of web based lectures. In addition to screening the materials, educators serve an important function in reaching out to large masses of people. By targeting twenty consumers, the lecture can educate twenty people. By targeting twenty educators (or mentoring the mentor), the lecture is reaching the educators and all of their students.

2.3. Quality control of biomedical information on the Internet Components of quality control

Although the "quality" term is used when we think about an excellent product or service, the same concept can be utilized to evaluate the performance of Internet based lectures. Experts generally recognize several distinct dimensions of quality that vary in importance depending on the context in which a quality control effort takes place. Epidemiological lectures also represent a product, thus we should be able to measure their quality as well.

Quality can be quantified as follows:

Q=P/E

Where Q= quality P = Performance E = Expectations

Quality control is the use of techniques and activities to achieve, sustain and improve the quality of a product or service. Garvin identified nine dimensions of quality, described in Table 1. The third column describes how quality dimensions identified by Garvin can be related to the quality of the Internet based lectures.

Table 1:   The Dimensions of quality

Dimension	Meaning and Example	Applications for Internet based lectures
Performance	Primary product characteristics, such as the brightness of the picture	Primary lecture characteristics, such as content, presentation, and relevance
Features	Secondary characteristics, added features, such as remote control	Secondary lecture characteristics, such as presence of references or sounds
Conformance	Meeting specifications or industry standards, workmanship	Meeting the standards of epidemiology education curricula
Reliability	Consistency of performance over time, average time for the unit to fail	The amount of time that lecture remains up to date
Durability	Useful life, includes repair	Useful life, includes updates
Service	Resolution problems and complaints, ease of repair	Resolution problems and complains, ease of lecture

		update
Response	Human-to-human interface, such as the courtesy of the dealer	Human-to-human interface, contact with the lecture developers or coordinators of Internet based libraries
Aesthetics	Secondary characteristics, such as exterior finish	Secondary characteristics, such as pretty web design
Reputation	Past performance and other intangibles, such as being ranked first	Being ranked first, having links from top institutions, having top scientists donating and using lectures

Adapted from David A Garvin *Managing Quality: The strategic and Competitive Edge* (New York: Free Press, 1988)

In the industry, performance is one of the major factors in determining the overall productivity of a system. Performance may include primary product characteristics, such as the brightness of the picture. Performance can also be measured for Internet lectures through review forms. In the context of the Supercourse, performance can be assessed by looking at the "overall" score of the lecture. Features are the secondary characteristics of the product, such as a remote control for TV or the presence of sound effects in a lecture. Conformance is an affirmative indication or judgment that a product or service has met the requirements of a relevant specification, contract, or regulation. For epidemiological lectures, conformance can be defined as adherence to the standards of epidemiology education curricula.

Reliability is quality over the long run. Reliability is the probability that a product will perform its intended function satisfactorily for a prescribed life under certain stated environmental conditions. From the definition, there are four factors associated with reliability: (1) numerical value, (2) intended function, (3) life, and (4) environmental conditions.

In technology, durability can be defined as the ability to exist for a long time without significant deterioration. The concept of durability is also relevant to epidemiological lectures.

For any educator in the area of public health it is very important to know whether the lecture on the web is up to date and whether it needs any updates.

Service is the ease of product repair, and response is the human-to-human interface. Both service and response may pose potential problems for the Internet based lectures. Due to the growing and ever evolving nature of the Internet, timely lecture updating could be problematic. Aesthetics refers to secondary product characteristics, such as the exterior finish of a product. Pretty web design is an aesthetic element of a lecture, which may also influence the quality judgment of a lecture reviewer.

Finally, reputation is the strategic standing of the organization in the eyes of its customers. Reputation is the "good name" of an organization resulting from its past performance. Good reputation has been at the core of the quality control mechanisms of the Supercourse lecture library. Lectures from top institutions such as Harvard add to the quality of the Supercourse collection. Similarly, lectures from Nobel Prize laureates add a valuable dimension to the quality of the Supercourse.

Although there is no literature on the quality control of web based epidemiological lectures targeting educators, there are studies addressing quality control of health data on the Internet targeting consumers. We reviewed a large number of studies in this area to obtain a better insight as to the general evaluation criteria for health related websites. These papers provided an excellent overview of the criteria commonly used for the evaluation of web sites, including credibility of authors, credibility of institutions, timely updating of the information, etc. Although tangentially related, they do provide the insight needed to identify evaluation criteria for our study.

Our literature review revealed that research studies in the area of quality control of the materials on the Internet target three major evaluation areas: structural measures of quality, performance measures, and user surveys. Several studies aimed to incorporate several evaluation areas into a single tool (Seidman et al. 2003). Clearly, more than one approach is needed to obtain the best measurement of quality of web site content.

2.3.1. Structural evaluation studies

Structural evaluation measures traditionally address the underlying systems and infrastructure, whereas process measures assess the extent to which health care providers have done the right things. Structural characteristics include aspects such as explanation of methods, validity of methods, and currency of information (Seidman 2003). Structural evaluation studies also examine content generation, credibility of authors, and updating process, without looking into the content itself (see table 2). Such studies mainly look at the websites and search engines providing patient oriented information, such as updates for breast cancer (Hoffman-Goetz et al. 2000).

Table 2: Structural evaluation studies: methodology and measurement

Methodology	Measurements
I. Explanation of methods	Explanation of content generation Listings of authors' affiliations Credibility of institutions Contact information for developers and interactivity
II. Validity of methods	Assertions supported by referenced material Peer reviewed content
III. Currency of information	Timely updating Date of the last update Site creation date

One of the common assumptions of structural evaluation studies is that if materials are coming from credible institutions and/or authors, have references, and are updated regularly, such materials are most likely to be of high quality. The idea is that the quality of materials coming from a professor from Harvard is higher than the quality of materials coming from an instructor from the Community College of Allegheny County. Peer reviewed content (whatever that may be) is valued more than something that is not peer reviewed. Regularly updated materials are viewed as better in quality, compared to those that are not updated. The structural evaluation model is efficient because it can provide a rapid assessment of the quality of the website by using a set of simple tools. A big disadvantage of this approach is however, the lack of consideration for the content of the website, as even professors from Harvard may post web materials of poor quality.

2.3.2. Performance measurement

Performance measures of quality evaluation on the web are concerned with the quality and comprehensiveness of information itself, not just the credibility of the source. For example, a study undertaking performance measurement of a diabetes education website would look at the presence of various criteria related to diabetes care (differentiation between various types of diabetes, diabetes testing, treatment options, etc) (Seidman 2002). The concept of performance measurement was highly relevant to this study because we were trying to look at multiple components of quality measurement, including lecture content

Table 3: Process evaluation studies: methodology and measurement

Methodology	Measurements
I Comprehensiveness of information	Presence of complete listings of aspects related to a certain disease and its control (screening, tests, immunizations, medications, etc)
II Accuracy of information	Presence of up to date information about the condition (disease specific) Comparison to existing criteria

Performance measurement studies in the area of Internet and health look at the comprehensiveness of information and the accuracy of information, based on the criteria established in a certain medical field (e.g. diabetes). The main strength of these studies is the effort they make to take an objective look at the content and content generation of health related websites. The problem with this approach is that as you move from one country to another, health recommendations may differ. It may not be appropriate to judge the comprehensiveness of immunizations site in Russia, based on the evaluation criteria coming from the US.

2.3.3. Consumer surveys/consumer feedback

Quality evaluation studies utilizing consumer survey research often rely on convenience samples of patients using the Internet for health information. Although patients oftentimes find Internet to be a valuable source of health related information, research suggested that Web sites need to be evaluated to ensure that the information they provide is accurate and current (Oermann et al. 2002). Consumer surveys on the Internet are often presented in the form of Likert scale questions, the format often used for the evaluation of educational materials.

Another measurement of consumer satisfaction with the website quality cited in the literature is the number of hits a certain website generates. Search engines like Google have

successfully explored the idea that quality may be a function of utility. The concept is that if a certain website is used more often than another, and if other websites link to this website frequently, the quality of this website may be better (this concept is described in more details in the section about the Google search engine). Thus, links to and from certain materials on the Internet may provide an interesting insight for the quality control.

Google Scholar enables you to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research. Just as with Google Web Search, Google Scholar orders your search results by how relevant they are to your query, so the most useful references should appear at the top of the page. This relevance ranking takes into account the full text of each article as well as the article's author, the publication in which the article appeared and how often it has been cited in scholarly literature (http://scholar.google.com/scholar/about.html#about). Just like in the citation index for biomedical literature, the number of citations or links from the authoritative sources provides important information about quality in the Internet based materials.

2.4. Web Based Peer Review

For many years, web based peer review has been an important issue discussed by the editors of web based journals. The approach that we utilize in the Supercourse is similar to the peer review system in a way that the lectures targeting educators end up being evaluated by peers. The dissimilarity of the approach is that what we do is not as formal as the traditional peer review process. Whereas the traditional peer review system is used to screen out certain materials from being published, our approach is used to enhance existing materials on the web and to use better and better retrieval systems to find the materials scientists need.

Traditional peer review has been thought to serve several purposes: communication of experimental results to scientific colleagues, precise record of methodology and findings, reduction of the likelihood that faulty data will be published, and establishment of scientific reputation (Editorial Incorporating the Internet. Nat Immunol. 2004 Jan;5(1):1). With the introduction of Internet publications, issues have been raised about the possibility of performing quality control entirely on the web.

The British Medical Journal (BMJ) was one of the first journals to become available on-line. Although BMJ is still a peer reviewed journal, it accepts correspondence letters (rapid responses) and makes them available on-line without peer review. BMJ search engine includes options that can do a search that includes or excludes rapid responses. It has been suggested that those who want to see the world as it is — rather than how they would like it to be — include rapid responses in their search (Smith 2004).

Although BMJ made rapid responses available without peer review to achieve democracy and balance, not everyone is enthusiastic about this effort (Delamothe 2002) "Recently the *BMJ* has published, alongside thoughtful letters from experienced scientists and clinicians, letters from the mad, the bad, and the misinformed" (Meadow 2002) said one of the authors who was criticizing the accuracy of rapid response letters from a clinical viewpoint. This raises a very important dilemma of whether "democracy" in the field of health related web information would work.

2.4.1 Automated Quality Control on the Web

Google.com-"democracy on the web"

Google is a play on the word googol, refers to the number represented by the numeral 1 followed by 100 zeros. It's a very colossal number: There isn't a googol of anything in the universe: not

stars, not dust particles, not atoms. Google's use of the term reflects the company's mission to organize the immense, seemingly infinite amount of information available on the web. Google examines more than 4 billion web pages to find the most relevant pages for any query and typically returns those results in less than half a second. Though a basic Google search answers most questions, it is possible to customize everything from the language of the interface to the format of the pages Google returns as results.

Quality control of Google search is carried out without human involvement or manipulation of results, which is why users have come to trust Google as a source of objective information untainted by paid placement. There are several mechanisms for Quality control of the Google search engine. The sites that are cited by other sites are given the priority as higher quality websites. Sites that are accessed more often (or have higher utility/usability) are also given a higher ranking.

The model used in Google is very similar to the model used for citation indexes. Citation indexes are used to find journal articles, and then determine where material has been cited and whether many writers in a field cite the work. Many scholars and editors use this type of index for quality control purposes, to determine whether or not a particular work has credibility (Lungberg 2003). Thus Google's system of quality control in a way mimics the criteria that have been used by scientists for many years. Objective systems of quality control like this one may also eliminate the problem of bias associated with conflict of interest, e.g. in drug companies. Since Google gives the user multiple articles on the same topic, the user may take a look at more than one article and/or obtain more information about the author and his or her affiliation.

2.5.   Peer review in Consumer Reports

The approach we are using is similar to the web based Consumer Reports model, as the rating scales we are using and the feedback process we have set up are very similar to major consumer reports websites. The reviewers are not necessarily experts, rather they are individuals who have experience with the "product". Consumer Reports guide thousands of users to the high quality products. The process is efficiently performed on the web, without much professional input and thus is very cost effective. Why can't we harness the experience of consumer reports websites to create something similar with health information? The major dissimilarity is the fact that instead of consumer products such as books or cars, our product is health information targeting a specific group of people. Nevertheless, Consumer Reports models provide a unique insight into the evaluation of web based information.

The common tool for measuring attitudes that is often utilized in consumer reports is Likert scales. The Likert technique presents a set of attitude statements. Subjects are asked to express agreement or disagreement on a five-point scale. Each degree of agreement is given a numerical value from one to five. Thus a total numerical value can be calculated from all the responses (Psychology of communications: attitudes

http://www.cultsock.ndirect.co.uk/MUHome/cshtml/index.html. Another tool that is often used in this field is the Semantic differential, a verbal rating scale that uses bipolar adjectives on a seven or five-point measure scale to measure beliefs, emotions, or feelings (Assael 1995). The Consumer Reports model relies on consumers interested in specific products. It does not provide a representative sample of consumers, yet people use Consumer Reports to obtain high quality reviews about the products they are interested in.

2.5.1. Amazon.com

Outside the field of public health, a very effective system of consumer based quality control has been generated by Amazon.com. Amazon.com was one of the first Web sites to allow public book reviews. In Amazon, any user is able to rate the book and view the ratings and comments of other users. Ratings, that are based on 5 point Likert like scales (1 being the worst and 5 being the best), are used to create an overall rating of the lecture: 1 star, 2 stars, etc.
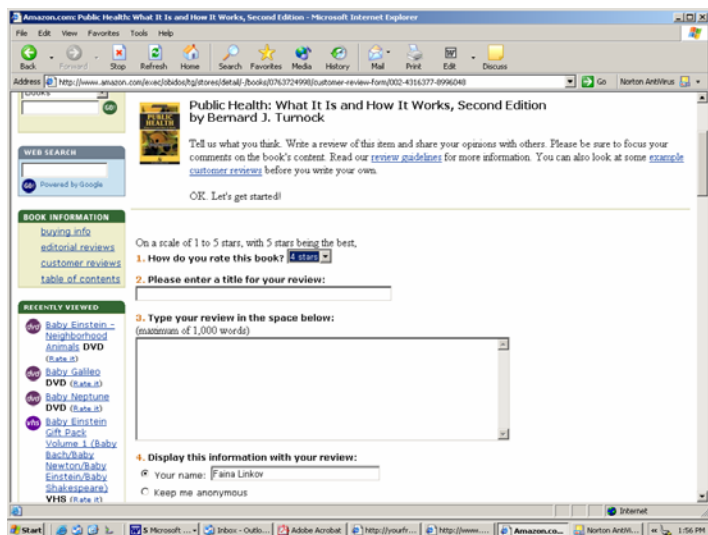
Figure 1:   Example of review page for Amazon.com

Over time, the reviewers have been divided into several categories: editorial reviewers (those associated with amazon.com), customer, and spotlight reviewers. A reviewer becomes a spotlight reviewer by a form of popularity test. At the end of each posted review, readers are asked to vote, "Was this content helpful to you?" Reviewers who receive a sufficient number of "yes" votes are promoted to the category of spotlight reviewer and their reviews are given prominence. Thus, Amazon is encouraging reviewers to provide helpful information in their feedback. Recently the approach towards the rating of Amazon.com products has been questioned due to tampering and misrepresentation (Harmon 2004), however it remains one of the most popular quality control systems for products purchased from the web.

We have suggested that the Amazon system could be used in the area of quality control for materials in the field of public health. Proposed research utilized user survey for the evaluation of the quality of public health lectures.

2.5.2 Epinions.com: "a web of trust"

Epinions.com, a website that provides opinions of customer products submitted by readers, introduced the concept of voting for reviewers. To encourage conscientious reviewing, the site has a complex process by which readers review the reviewers. Respected reviewers receive recognition, such as cash awards or having their photographs added to the Web site.

The Income Share program at Epinions.com rewards writers who contribute reviews that help other users make decisions. Epinions takes a share of the revenue gained from providing consumers with high-quality information and deposits it into good reviewer's account. Income Share bonuses are not tied directly to product purchases, but are based instead on more general use of reviews by consumers making decisions.

The staff of Epinions.com or outside consultants do not review comments submitted to Epinions.com. When you preview a review, the Epinions.com spell-checker and language filter will highlight any problematic words so that you can make your own changes. Once the review passes these automated checks and you publish it, it will be available to others. Epinions does not manually change the text of published reviews.

On Epinions.com the user can create "a web of trust". One's Web of Trust is a network of reviewers whose reviews and ratings one has consistently found to be valuable. The Web of Trust mimics the way people share word-of-mouth advice every day. For example, friends have a

proven track record. If a friend consistently gives you a good advice, you're likely to trust that person's suggestions in the future.

2.5.3 ConsumerReports.org

ConsumerReports.org provides yet another option for website evaluation by generating specific evaluation criteria. It evaluates the credibility, usability and content of shopping, service, and information web sites. In ConsumerReports.org e-Ratings**,** the Overall score is a reflection of the evaluation of a site's credibility, usability, and content, and how these components come together to create a satisfying, efficient, and effective online experience. The Credibility score reflects the quality and clarity of a site's explanation of privacy, security, and customer service policies, and the disclosure of pertinent business-related information. The criteria for this score were developed by Consumer WebWatch and ConsumerReports.org**,** based on Consumer WebWatch's guidelines for improving web sites. The usability score reflects the ease and efficiency with which a site can browsed and searched, as well as the ease of placing an order. The content score reflects the breadth and depth of product and information categories and choices within those categories; the amount and quality of information available; and the availability of useful personalized/customized, special, or unique features.

2.5.4 Angieslist.com

Angieslist.com - "ask your neighbor" solution to quality control of homeowners' services

In 1995, a woman named Angie Hicks became concerned about low quality services offered to homeowners. She, together with her friends and neighbors, started a list of good and

bad service companies. Every time one of them hired a company, they told Angie how they did. Angie's List became the only source of independent, unbiased service ratings in the city.

Today, Angie's List is active in fifteen major markets and has ratings on more than 10,000 service companies. More than 100,000 homeowners use Angie's List to find good service in 250 categories, including roofing, plumbing, landscaping and auto repair. Membership in angieslist.com is providing homeowners with satisfaction ratings of thousands of homeowners across the US.

In addition to the consumer reports options described above, there are many other web and product evaluation options out there. For example, CNet.com provides information for those who would like to purchase a computer. In addition to product ratings, CNet provides information about the places where computer can be purchased and the ratings of the stores. Many things can be learned from consumer reports and many aspects of the existing system can be easily incorporated into the evaluation of biomedical literature.

In general, web based approaches to quality control are aiming to speed up and automate the process of quality control to provide "actionable" information to readers. The major strength of all these approaches is high speed and high throughput. The cost of evaluating large numbers of materials is reduced dramatically with the majority of web approaches. Another positive aspect of these systems is the fact that they not only aim to evaluate the quality of the product, but also the quality of the reviewer. Assessment of review quality is something that is not well addressed in the peer reviewed biomedical journals. Articles are often assigned to reviewers with very limited knowledge and expertise in the area they are asked to review. The most obvious disadvantage of these approaches is vulnerability to tampering and abuse, as demonstrated with amazon.com. It is anticipated that with further development of web based quality control

systems, their susceptibility to abuse will be reduced. Consumer satisfaction is expected to remain one of the key criteria for web based quality control systems; however there have been suggestions made about modifying the eligibility criteria for posting the reviews.

2.6.   Educational Program Evaluation

Evaluation of teaching, teacher effectiveness, and teaching materials has existed as long as there has been teaching. Generations born two thousand years after Jesus and Socrates still evaluate the teaching of these masters (Beecher 1949). Evaluation of teaching materials is at the core of this dissertation; therefore we decided to look into the field of education and educational program evaluation to get a better insight into common evaluation tools.

Evaluation of teaching materials is closely tied to the evaluation of teaching, which became a very popular topic of research in the U.S. in 1930's and 40's when school teaching became a reputable job. In 1932, Renis Likert invented a measurement method, called the Likert Scale, that is currently used in attitude surveys. These scales allowed answers that ranged from "strongly disagree" to "strongly agree" and became very popular in the field of teaching evaluation. The project presented in this dissertation research utilized Likert-like scales to develop a quality control tool for the Internet based lecture library.

It is hard to find an ideal way to evaluate the educational program, as each method has its own advantages and disadvantages and virtually no standardization. That is why the methods are often combined. Common types of research used for program evaluation include descriptive study, relational study, and experimental or quasi-experimental research (Ary 1985). Evaluation may involve subjective and objective measures and qualitative and quantitative approaches. The resources devoted to evaluation should reflect its importance, but excessive data collection

should be avoided. A good system should be easy to administer and utilize information that is readily available. (Morrison 2003)

In addition to the field of teaching evaluation, we also looked into the field of training effectiveness, commonly used for the evaluation of professional training programs. Effectiveness often entails using the four-level model developed by Donald Kirkpatrick. According to this model, evaluation should always begin with level one, and then, as time and budget allows, should move sequentially through levels two, three, and four. Information from each prior level serves as a base for the next level's evaluation (Kirkpatrick 1994)

Figure 2:   4-level model developed by Donald Kirkpatrick

Focus groups are commonly utilized to narrow down the scope of the evaluation studies by targeting the priority areas. Questionnaires and surveys are very common tools utilized in program evaluation research. Proposed effort was a program evaluation project in the area that has been rarely researched before: Internet education program targeting teachers. Detailed description of the program is included in the next section.

There are a very limited number of studies evaluating Internet based teaching materials. One of these studies, conducted by Zhang in 2003 looked at the evaluation of a distance learning course conducted via Blackboard. Lecture evaluation tools this project utilized included the use of questionnaires and line usage. Usage was measured by looking at the total number of accesses

per website, number of accesses over time, user access per hour of the day/per day of the week, and total accesses by user.

Another project aiming to evaluate a distance learning course took place in the University of Sussex. The following criteria were proposed to evaluate the lectures: student feedback (based on the Likert style questionnaire and focus group), lecturer feedback, and student usage of the lecture. (http://www.sussex.ac.uk/press_office/bulletin/22feb02/article9.shtml)

In 1998, Oliver and Conole surveyed means of evaluating communication and information technologies (C&IT). Tools available include evaluating on line usage, and use of questionnaires. Thus, we found many similarities in the evaluation tools in the field of training evaluation, lecture evaluation, teaching evaluation, and internet materials evaluation.

2.7.   Global Health Network Supercourse project

2.7.1.   Background and current status

The Global Health Network Supercourse Project started in the fall of 1997 (Aaron et al. 1999). The Supercourse is the library of lectures on prevention, epidemiology, and global health. It is currently available at www.pitt.edu/~super1 and anyone can access it free of charge. Supercourse is not a course by itself but is a collection of independent lectures written by authors who want to share their experience with people of other countries (Acosta et al. 1999). During the initial development of this effort, the Supercourse was funded for three years by a grant from NASA and it is currently funded by the National Library of Medicine. The name and the idea of the project evolved as the result of Drs. LaPorte, and Songer teaching a class in chronic disease epidemiology in 1995. To their knowledge, this was one of the first international efforts to provide Internet training.

Teaching the Teachers: The Supercourse is not a substitute for existing educational systems, but a teaching-support system. It provides high level lectures to the teachers of students in medical, dental, nursing schools, and those of public health etc. These are passionate lectures by experts in the field, and the teacher just "takes" them out like a library book to teach. The Supercourse is not just a distance education model for two reasons: The first is that despite our effort being global, there is a "death to distance" as the Economist has quoted. This means that if a student is in the next room, or in the next continent, it makes no difference. In addition, distance education means a separation between the teacher and the student. Here we have no separation in that the classroom teacher are doing the teaching, but they will have much better educational lectures than they ever had before.

As of March 10, 2005, the Supercourse contained over 2156 lectures donated by over 20,300 members of the Global Health Network from over 151 countries. The Supercourse project inspired a variety of Internet based networks such as: Islamic network, Indian Heritage, Women's Health, Former Soviet Union, Pakistani network, and many others. The Supercourse has a variety of multilingual lectures translated into 13 languages. One of the major directions of Supercourse development is the evaluation of quality of the Supercourse lectures.

Figure 3:   Front page of the Global Health Network Supercourse Project.

With over 20,300 participants world wide, the front page of the Supercourse website receives thousands of hits monthly.

2.7.2. Presentation format and review forms

Supercourse lectures are transformed from "traditional" PowerPoint presentations into condensed format. In the past decade, PowerPoint became "a language of science" that could be easily understood by faculty members all over the world. Many faculty members use PowerPoint presentations to present the information about their research to their students and colleagues. Many speakers, instructors, and faculty members put their PowerPoint presentations on the web to make them publicly available.

PowerPoint presentations are large in terms of their size. The average size of PowerPoint presentations ranges anywhere from 0.5 to 10 megabytes and these numbers can go even higher if these presentations have graphics. It might take up to one hour to download a 5 megabite presentation with a regular modem for someone who is located in the United States. The situation becomes even more difficult for someone who is located outside of the United States, especially in a developing country. Supercourse format was designed to condense PowerPoint format into something that would be accessible for those with limited Internet access.

Each Supercourse lecture consists of 14 to 32 consecutive pages and every page has a uniform format: a slide with 320 by 240 pixels in size on the left and text beside the slide on the right. On the last page of each lecture is the peer review form for the lecture. This page allows the readers of the lecture to rate and give comments on the lecture (Sekikawa et al. 2000). Review forms of the Supercourse lectures became the basis of quality control for the Supercourse lectures and the data collected through these forms helped us to test our hypothesis.

The lecture review forms (see figure 4) have the following questions: name, position, organization, e-mail address, have you ever taught and introductory epidemiology course (yes or no), rate the lecture on content, presentation, relevance, and overall rating (a rating scale for these four items: 5=excellent, 4=above average, 3=average, 2=below average, 1=poor). In May 2004, Dr. Songer suggested that the "expectation" rating of the lecture may provide many important insights into quality measurement in the Supercourse. In August 2004, additional question was added to all of the Supercourse peer review forms: "How does the quality of the lecture compare with your expectations about it?" (The rating scale for this item also utilizes Likert scales: 5= Well above what I expected, 4= Above what I expected, 3= Same as expected, 2= Somewhat below what I expected, 1= Well below what I expected.)

Figure 4: Lecture review form, Supercourse project

2.7.3. Quality Control of the Supercourse lectures

The Supercourse ensures quality of lectures in more than one way. One of the most obvious quality control mechanisms has to do with structural measures of quality. Supercourse developers make sure that the lecture comes from a trustworthy source. We have lectures from the leading schools in the US and worldwide including Harvard, Johns Hopkins, etc. Cutting

edge researchers, including 6 Nobel Prize laureates and 5 heads of NIH have also contributed their lectures.

Since the audience of the Supercourse is not composed of healthcare consumers, but of educators, we are adding another level of quality control: educators can judge the quality of a lecture and update it based on their research findings, cultural specifics, and/or geographical location.

Consumer Education Model vs. Faculty Member Education Model

Information
Repository

Consumer with access to
information repository

*quality control is
limited

Information
Repository

Teacher/instructor with
access to the information
repository

consumer

*quality control is taking
place on the instructor
level
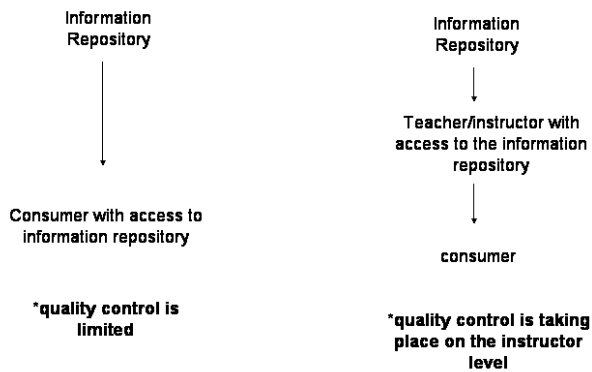
Figure 5: Comparison of 2 Education models

The focus of this project was the research of open peer review system of the Supercourse reviews. Just like in the traditional peer review model, there is more than one way to look at the quality of a Supercourse lecture. Quality may be looked at as a function of content or as a function of presentation/lecture delivery. When looking at the lecture delivery, someone may rate a lecture as poor if it takes a long time to download or if the information is presented in a boring way. For the purpose of this research, we focused on the quality of the content, and not on the lecture delivery.

3 METHODS

3.1 Comparison of Supercourse lecture reviews to the reviews of experts

One of the main goals of this study was to compare Internet based evaluations to the "Golden standard" evaluations provided by Epidemiology experts. We wanted to test the hypothesis that there is no statistically significant difference between the mean overall scores of Internet reviewers and epidemiology experts. In order for us to test this hypothesis we needed to identify a random sample of Supercourse lectures that would undergo expert reviews, select expert reviewers, and collect the data.

3.1.1 Lecture selection

Sample size formula for correlations has been utilized to estimate the sample size for this research. We found that in order to detect a correlation coefficient of 0.7 (a large correlation), 100 lectures would need to be selected for the sample.

As of today, 2156 lectures have been a part of the quality control process. Of the first 1000 lectures we received 849 reviews, with at least 250 lectures accumulating at least three reviews each. It was decided to concentrate our evaluation efforts on the first 1000 lectures, since those were the ones that accumulated the maximum number of reviews. All lectures in foreign languages were excluded from this research. Lectures that had multiple parts were evaluated as one lecture. After excluding foreign language lectures, we had a set of about 200 lectures that had three reviews and more. One hundred lectures were randomly picked out from this set using computer generated random numbers.

One of the possible problems associated with selecting lectures that accumulated three or more reviews is the fact that there may be something different about those lectures. Lectures that undergo more evaluations may potentially be of higher quality. In addition, lectures that accumulated more reviews are also older lectures. In the Supercourse, lectures that were accumulated when the project was first launched in the late 1990's came only from the top faculty members in the field of epidemiology. Again, the quality of these lectures may be higher. Additionally, the developers of the Supercourse encouraged members of the network to review new lectures when the project first started, however this practice diminished as the library grew in size.

We performed a small pilot study to figure out how long it takes to review one lecture. Several Supercourse collaborators reviewed five lectures and gave us the estimates of the time it took them to complete the reviews for these lectures. We estimated that it takes six to fourteen minutes to complete one lecture with slightly higher estimates for non-English speakers. Thus, we estimated that in order to evaluate the entire set of 100 lectures, an expert would need to spend at least seventeen hours. The total of all hours donated by all reviewers was large: at least 119 hours.

3.1.2 Selection of experts

The research study was advertised through the newsletter of the Global Health Network Supercourse project. Sixteen people responded to the letter and expressed interest in participation. Nine of them were either unable to donate the required time or did not fulfill the research participation criteria. A big advantage of our recruitment approach is that the experts that we ended up selecting were both, experienced and well published researchers and they were

the end users of the Supercourse. The ratings of our experts thus combined the expertise of the researchers with the expectations and behaviors of the Supercourse user.

Seven experts from 6 countries agreed to participate in the project. International reviewers were targeted because we wanted to better mimic the current system of peer review in the biomedical journals. With the globalization of science, more and more international scientists are asked to serve as the reviewers in the biomedical journals. The summary of our experts' qualifications is presented in the table below. "Expert" has been defined as someone with a PhD in the area of epidemiology, research experience, and evidence of scientific publications in peer reviewed journals. All of the six volunteers possessed the necessary experience to serve as expert reviewers. Three of the experts currently serve as editors for major biomedical journals.

Table 4: Summary of Expert qualifications

Reviewer identifier	Country	Degree(s)	# papers published	Peer review experience
Reviewer1	UK	MD, MPH	3	Local*
Reviewer2	India	MBBS	6	7 papers reviewed
Reviewer3	France	MD, PhD	60	Editor of Angiology journal
Reviewer4	USA	DDS	10	Editor, Dental journal
Reviewer5	USA	MPH, DSc	50+	Editor of 5 journals
Reviewer6	China	PhD	18	Local
Reviewer7	Cuba	MD, MPH	Local	Local

3.1.3 Information collection: website development

We created a website dedicated to this research project (see figure 6). The front page had all the instructions for expert lecture reviewers. It also had a link to all 100 lectures. All instructions were also confirmed with the research participants via phone and/or e-mail. Six research participants were given 2 months to complete the project.
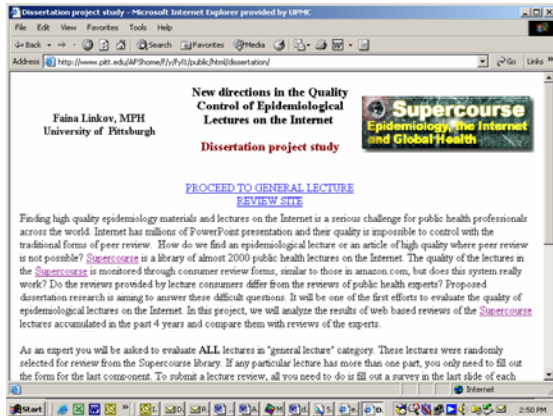
Figure 6:   Front page of the Website Developed for this Research Project.

3.1.4   Exploration of descriptive statistics

Upon the completion of data collection for the Supercourse (non-expert) and expert reviewers, we wanted to explore the descriptive statistics of these data. We hypothesized that the lectures would be viewed positively by both experts and non-experts, with the majority of scores ranging between 3 and 5. For looking at the descriptive statistics, we will calculate the means and standard deviations for the scores of expert and non-expert reviewers. We also looked at data ranges of individual reviewers. Box plot of the data was constructed to ease the evaluation of basic score distribution.

3.2   Testing the difference between the means

Two sample t-test was performed to see whether there was a difference between the mean overall scores of experts and internet reviewers. Two sample t-test was also utilized to see if there was a difference between the mean overall score of experts who are editors and the mean overall score of experts who are not editors.

We also looked at experts' personal backgrounds to see if they may play a role in reviewers' scoring pattern. We collected CVs/Resumes of all 7 experts, thus it was easy for us to

identify which lecture in the set of lectures we selected matches their areas of expertise and may be of interest to them. For each reviewer, we went through his or her overall scores and separated them into 2 categories: interesting (coded at 1) and non-interesting (coded as 0). Interesting lecture was defined as a lecture that came from the field of the reviewer's expertise. For example, for the reviewer who was the expert in cardiovascular health, all cardiovascular lectures were marked as "interesting". Non-interesting lectures referred to lectures that were outside of the scope of interest of the reviewer. We utilized t-test to see if there was a difference between the mean score of "interesting" vs. "non-interesting" sets lectures for each reviewer. The same procedure was utilized for all 7 reviewers.

3.3 HALO effect

The history of Halo effect research started in 1920's with the groundbreaking work of Edward Thorndike. Edward Thorndike found that when army officers were asked to rate their charges in terms of intelligence, physique, leadership and character, there was a high cross-correlation among the ratings (Thorndike 1920). Thorndike's research suggested that when we consider a person good (or bad) in one category, we are likely to make a similar evaluation in other categories. Halo effect has also been defined as a systematic bias in attribute ratings resulting from raters' tendency to rely on global affect rather than carefully discriminating among conceptually distinct and potentially independent brand attributes. (Leuthesser et al. 1995) This tendency towards consistency manifests itself as higher-than-actual correlations between attribute ratings because individuals are psychologically motivated to "level out" discrepancies which appear in belief structures at a micro level (Beckwith et al, 1978). In general and outside the field of psychology, Halo effect occurs when good or bad performance in

one area affects the assessor's judgment in other areas and 'leniency' (Anastasi 1982). We hypothesized that we would be able to observe halo effect in the Supercourse ratings.

Approaches to measuring the halo effect have ranged from simple observance of the average inter-attribute correlations to factor analysis of the rating data coupled with statistical correction for halo. Although it is difficult to state with any degree of precision the point at which halo is present, a rough rule of thumb is that average inter-correlations of around 0.60- 0.70 or greater are suggestive of a halo effect (Leuthesser et al. 1995). In this research, we implemented the basic approach to measuring halo by constructing correlation matrix.

Overall, presentation, relevance, content, and expectation scores are regularly collected for all Supercourse lectures. We hypothesized that just one score (the overall score) may be sufficient to judge the quality of the lectures, as we suspected that the correlation between the overall score and other scores would be high. We decided to calculate the correlation coefficient between the overall score and other scores to determine whether overall score was a good predictor of other scores.

3.4   Inter rater agreement

We hypothesized that the agreement among the reviewers would be high. Kappa statistics were calculated in order to look at agreement among the reviewers. Intraclass correlations were also calculated to analyze the similarities among the ratings. Intraclass correlation is ANOVA-based type of correlation that measures the relative homogeneity within groups in ratio to the total variation and is used. Intraclass correlation is commonly used to measure inter rater agreement. All statistical procedures for this dissertation were performed in SAS software package.

The Kappa statistic is a commonly used measure of agreement, or repeatability in epidemiological studies. Through the assessment of repeatability, epidemiologists can assess inter and intra-observer reliability of different procedures or instruments. Research suggests that at least 5 statistical packages: Stata, Systat, SAS, BMDP, and SPSS each compute Kappa correctly (Kim 2001). In this research, we are utilizing Kappa statistic to look at the agreement among the 7 reviewers. One way ANOVA was utilized to look at the agreement among 7 expert reviewers and the Supercourse reviewers.

The reason why both Kappa and ANOVA were used is because we could not include the Supercourse reviewers in the Kappa calculation. They could not be included because we did not have a fixed number of the Supercourse reviewers per lecture.

Quality scores or "Q" were be calculated for each lecture. We utilized the following formula Quality =Performance/Expectation, adapted from Besterfield, 2001. The formula suggests that if performance exceeds expectations, the quality would be high. Overall scores were used as a performance score. ANOVA was used to analyze the difference between the quality scores of 7 expert reviewers. We were not be able to calculate "Q" scores for non-expert reviewers because the question assessing the expectations was added to the Supercourse review form only in 2004. This means that older reviews were not able to accumulate enough expectation scores.

3.5 Exploration of quality criteria: Follow up with the experts

Upon the completion of the study, each expert was sent a follow up question:

"What kind of criteria did you use to rate lectures positively or negatively?

(please, list a few things that helped you in the lecture rating, such as research design, overall lecture flow, grammar, etc)"

The purpose of this question was to explore what kind of criteria the experts were using to judge the quality of the Supercourse lectures.

We hypothesized that we would observe high correlation between the overall score and other scores (content, relevance, presentation, and expectation) for both experts and non-experts, thus demonstrating the HALO effect. This effect is important because it may suggest that assessment of 1 score, overall score in this case, may be a good predictor of other scores. Correlation coefficients were calculated between overall scores and other scores using SAS software.

3.6 Personal Background: quality predictor?

We hypothesized that the reviewer's personal background may play a role in the way he or she scores the lectures. For example, if one reviewer's background is cardiovascular health, this reviewer may assign lectures in this area lower or higher score because of increased sensitivity to the content. For each reviewer, we decided to compare the mean of lecture scores for the area of this reviewer's interest vs. the lectures outside of his or her scope. "Interesting" lectures will be selected with the help of reviewers' CVs and resumes that were submitted to us prior to the beginning of this study. T-test will be utilized to analyze the difference between these scores for each individual reviewer.

3.7 Highly rated lectures and lectures that obtained low scores: Exploration of lecture characteristics

We hypothesized that there may be something interesting about the lectures that receive high scores vs. low scores. For example, we hypothesized that poorly rated lectures may have poor structure; lack of scientific references, and may be coming from unknown authors. We randomly selected 2 reviewers: one from a developing and one from a developed country. We

will go back to their overall lecture scores and select 10 lectures that were rated highest (score of 5) and lectures that were rated lowest (scores of 1-2). We will look at the characteristics of these lectures and will try to identify any interesting trends if such exist.

4. RESULTS

4.1. Descriptive statistics

All the data were collected over the period of 2 months and entered into an excel database. SAS was utilized for the statistical analysis of the data and Excel was used to create the graphs. The total of 658 lecture reviews were collected from 7 experts. We collected a total of 849 reviews from non-expert/Supercourse reviewers from the Supercourse websites, but the exact number of non-expert reviewers is hard to determine because some people choose to fill out the forms only partially. Detailed information about the descriptive statistic is summarized in the table and box plot. Overall, the lectures were reviewed positively by both experts and non-experts.

Table 5: Descriptive Statistics

Statistic	Expert (N=7) (based on 658 reviews)	Supercourse Reviewer based on approximately 849 reviews
Mean overall score	3.92	4.12
SD	0.95	0.82
Range	1-5	1-5
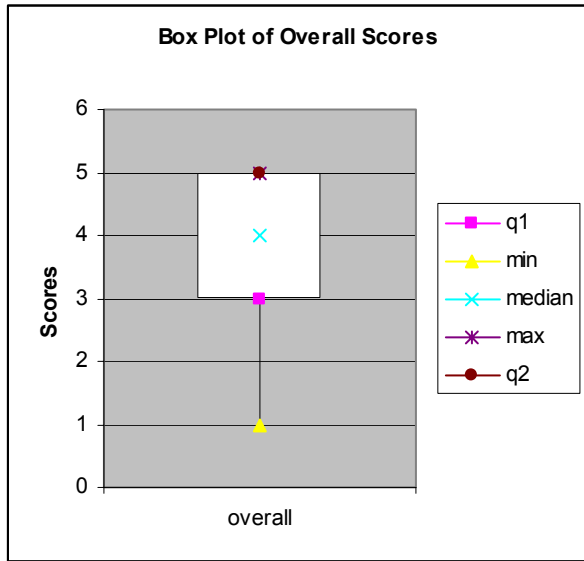
Figure 7: **Box Plot: Overall lecture score distribution**

This box plot provides an excellent visual summary of many important aspects of the lecture score distribution. It nicely demonstrates that at least 50% of the lectures were given a score of 3 and higher with the mean and the median score of 4.

By looking at the descriptive statistics for the ratings of the individual reviewers, one can observe that the reviewers' opinions on the lecture quality differ substantially: the means and the ranges of scores do not look consistent among the majority of the reviewers. Basic statistics for individual reviewers are summarized in table below.

Table 6: Descriptive statistics for individual expert reviewers and Supercourse reviewers

Reviewers		Rev1	Rev2	Rev3	Rev4	Rev5	Rev6	Rev7	Super
N		94	103	81	99	97	94	91	849
Overall	Mean	4.04	3.12	3.75	4.07	3.93	3.94	4.66	4.12
	SD	0.83	0.95	0.92	0.82	1.00	0.73	0.60	0.82
	Range	2-5	1-5	1-5	2-5	1-5	2-5	3-5	1-5
Content	Mean	4.11	3.14	3.79	4.38	4.25	3.92	4.78	4.18
	SD	0.76	9.94	1.02	0.77	0.84	0.63	0.51	0.74
	Range	2-5	1-5	1-5	2-5	1-5	3-5	3-5	2-5
Presentation	Mean	4.06	3.22	3.43	3.78	4.06	4.00	4.52	4.02
	SD	0.83	0.97	1.09	1.03	0.98	0.78	0.75	0.85
	Range	2-5	1-5	1-5	2-5	1-5	2-5	2-5	1-5

Relevance	Mean	3.98	3.12	3.69	4.15	3.71	3.94	4.67	4.34
	SD	0.82	0.96	0.87	0.85	1.21	0.77	0.68	0.73
	Range	2-5	1-5	1-5	2-5	1-5	2-5	2-5	1-5
Expectations	Mean	3.10	2.63	3.51	3.38	3.75	3.09	4.23	N/A
	SD	0.53	1.03	0.99	0.90	1.02	0.94	0.84	
	Range	2-4	1-5	1-5	1-5	1-5	1-5	1-5	

One way ANOVA was utilized to see if there was a statistically significant difference among the means of 7 expert reviewers and the Supercourse reviewers. Null hypothesis was rejected. The means were statistically different from each other (F=27.65 P<0.0001)

Bar charts below show the frequency distribution of expert and Supercourse (non-expert reviews. By looking at these figures, we can see that most reviewers prefer to assign the scores of 4 and 5, suggesting that there is a digit preference.

Based on N = 658 lecture reviews
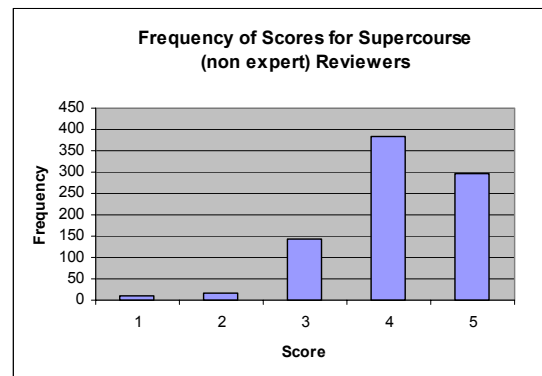
Based on N=849 lecture reviews

Figure 8:   Experts and non-experts: Bar chart of frequency distribution

4.2.   HALO effect

Through our lecture review form, we collected a total of 5 quality related scores for each lecture: Overall, presentation, relevance, content, and expectation score. We analyzed the correlation between the overall score and other scores in order to determine whether the overall score was a good predictor of other scores.

We found that the correlation between the overall scores and 4 other scores for experts was quite high ranging from 0.80 to 0.90. Due to the fact that the overall score was such a strong predictor of other scores, we continued our analysis by utilizing the overall score only. The strong correlation between the variables is graphically presented in the figure below.

	Overall	Content	Relevance	Presentation	Expectation
Overall		0.86	0.90	0.84	0.80

Overall score vs. Relevance Score — $R^2 = 0.8135$

Presentation Vs. Overall Score — $R^2 = 0.7111$

Content Vs. Overall Score — $R^2 = 0.7437$

Overall vs Expectation Score — $R^2 = 0.6186$

Spearman Correlation Coefficients: Prob > |r| under H0: Rho=0 (based on N=658 expert reviews)
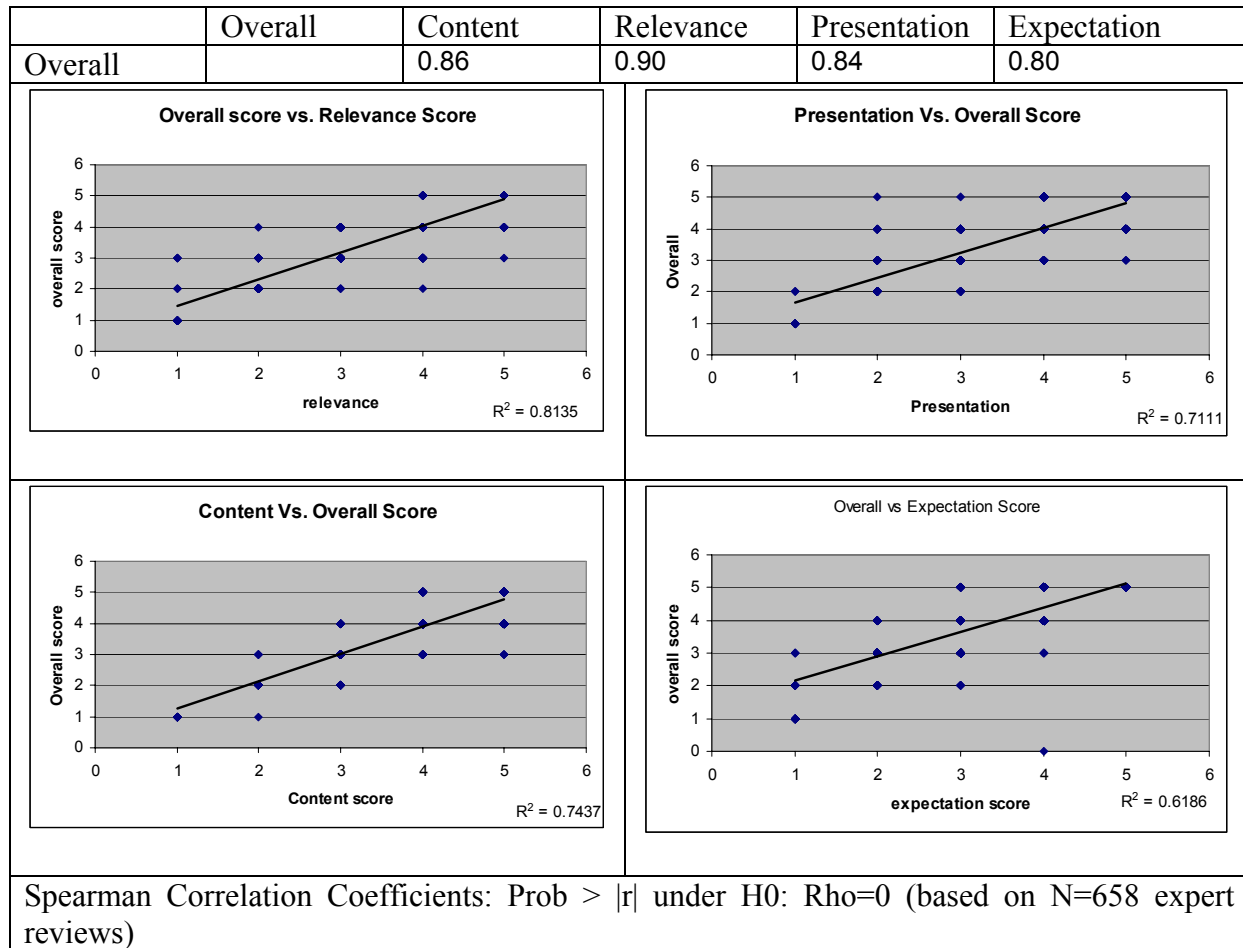
Figure 9: Visual demonstration of Halo effect for expert reviewers

Spearman correlation coefficients between content, presentation, relevance, and overall scores were calculated for non-experts (Supercourse reviewers as well). Non parametric correlation coefficient calculation was used since the scores were not normally distributed. Analysis showed that all scores were significantly correlated to each other. This indicated an apparent halo effect, where if one measure was viewed as positive, the others were as well. It

also suggested that we do not need 4 measures of quality, but rather only one, the Overall assessment.

Table 7:   Correlation among content, presentation, relevance, and overall score for non-expert (Supercourse reviewers)

	Overall	**Content**	**Presentation**	**Relevance**
Overall	1.00	0.78	0.71	0.70
Sig. level		<0.01	<0.01	<0.01
# observed	408	407	408	406
Content		1.00	0.53	0.62
Sig. level			<0.01	<0.01
#observed		407	407	406
Presentation			1.00	0.48
Sig. level				<0.01
#observed			408	406
Relevance				1.00
Sig. level				
#observed				406

Spearman Correlation Coefficients: Prob > |r| under H0: Rho=0

4.3.   Looking at the difference between the means

T-tests were utilized to compare the means of 2 samples (experts vs. non-experts) of lecture reviews. We compared the mean overall score of experts (calculated based on the total of 658 lecture reviews of 7 experts) with the mean of the Supercourse reviewers (non-experts) that was based on 849 lecture reviews collected over the past 4 years through the Supercourse website. Student T-test procedure was used in SAS software. Our results suggested that there was a statistically significant difference between the scores of experts and non-experts, with experts assigning lower scores (T=3.9, p <0.0002, null hypothesis of no difference rejected). The results of two sample t-test were confirmed by the non-parametric analogues. Although the results of t-test suggest that there is a statistically significant difference between the two means, this difference is not very meaningful because we had a very large sample size of lecture

reviews. Large size of review made our test very sensitive to detect even a small difference, despite the fact that the means appear to be very similar (3.9 vs. 4.1)

We also utilized two sample t-test to see if there was a difference between the mean overall score of editors who are editors and the mean overall score of experts who are not editors. We had a total of 3 experts who are editors who evaluated a total of 277 lectures; 4 non-editors evaluated 382 lectures. The mean score of non-editors was 3.91 (SD= 0.97); the mean score of experts who are editors was very similar: 3.82 (SD= 0.92). T-test indicated that there is no difference between the 2 means (t=-0.19, p<0.85)

4.4.   Inter rater agreement

Kappa statistic was calculated in order to look at the inter rater agreement among the 7 expert reviewers. Resulting Kappa suggested that inter rater agreement is very low for experts, consistent with the existing literature in the area of peer review. Kappa variables are presented in table 8.

Table 8:   Inter-rater agreement: Table Kappa statistics

	Reviewer1 N=94	Reviewer2 N=103	Reviewer3 N=81	Reviewer4 N=99	Reviewer5 N=97	Reviewer6 N=94	Reviewer7 N=91
Reviewer1 N=94	1	0.04	-0.03	0.06	-0.04	0.05	0.03
Reviewer2 N=103		1	0.04	0.02	0.01	0.02	-0.04
Reviewer3 N=81			1	-0.06	0.04	-0.01	0.04
Reviewer4 N=99				1	0.13	-0.05	-0.01
Reviewer5 N=97					1	0.12	0.01
Reviewer6 N=94						1	0.12
Reviewer7 N=91							1

Intraclass correlations were calculated to look at the inter rater agreement among experts and the Supercourse reviewers (non-experts). Resulting data suggests that experts' reviews poorly correlate among each other, as well as the Supercourse reviews. Intra class correlations are presented in table 9.

Table 9: Inter rater agreement: Intra class Correlation coefficients

	Reviewer2 N=103	Reviewer3 N=81	Reviewer4 N=99	Reviewer5 N=97	Reviewer6 N=94	Reviewer7 N=91	Sup. N=849
Reviewer1 N=94	0.49	-0.25	-0.45	-0.40	-0.43	0.07	-0.28
Reviewer2 N=103	1	0.31	0.12	0.19	0.03	0.24	-0.31
Reviewer3 N=81		1	0.17	0.14	0.12	0.12	-0.26
Reviewer4 N=99			1	-0.18	-0.33	-0.33	-0.17
Reviewer5 N=97				1	-0.33	-0.45	-0.38
Reviewer6 N=94					1	-0.84	-0.11
Reviewer7 N=91						1	-0.17

Intraclass correlation is large and positive when there is no variation within the groups, but group means differ. It will be at its largest negative value when group means are the same but there is great variation within groups. Its maximum value is 1.0, but its maximum negative value is (-1/(n-1)). A negative intraclass correlation is not common, but it occurred in our study. Negative intraclass correlations occurs when between-group variation is less than within-group variation, indicating some third (control) variable has introduced nonrandom effects on the different groups.

4.5. Quality

We calculated Q=Performance/Expectations for all lectures reviewed by the experts. The values were expected to range from 0.2 (poorest quality) to 5 (best quality). Q values above 1

would be considered as positive quality scores because that is when performance exceeds the expectations. The mean quality score for the expert reviewers was 1.22 SD 0.29. The table below demonstrates the mean quality score for each expert reviewer.

Reviewer	Mean	SD
Reviewer 1 N=94	1.31	0.20
Reviewer 2 N=103	1.26	0.40
Reviewer 3 N=81	1.11	0.23
Reviewer 4 N=99	1.26	0.28
Reviewer 5 N=97	1.07	0.18
Reviewer 6 N=94	1.35	0.29
Reviewer 7 N=91	1.14	0.25
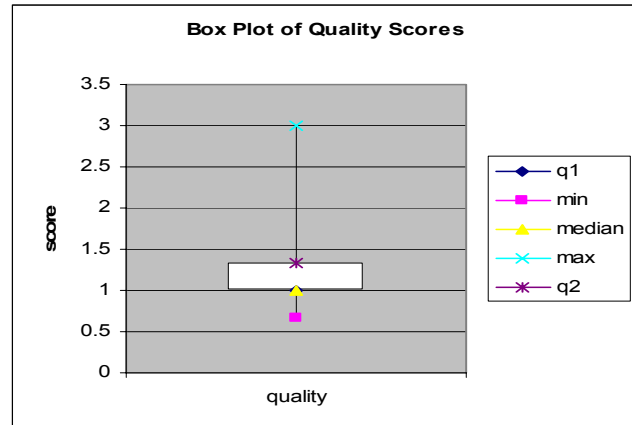
Figure 10: Quality Scores

ANOVA was utilized to see if these means are statistically different from each other. The test revealed that they are in fact statistically different from each other. F=14.12 P<0.0001

4.6. Expert's personal background as a factor predicting scoring pattern

We utilized t-test to see if there was a difference between the mean score of "interesting" vs. "non-interesting" lectures within the scores of each reviewer. Interesting lecture was defined as a lecture that came from the field of the reviewer's expertise. For example, for the reviewer who was the expert in cardiovascular health, all cardiovascular lectures were marked as "interesting". Non-interesting lectures referred to lectures that were outside of the scope of interest of the reviewer. The same t-test procedure was utilized for all 7 reviewers.

We found that for one of the experts personal background may have played a role in differential scoring, while for others it did not appear to make much difference. Reviewer #5 consistently rated lectures in his area higher than other lectures (t = -3.51, p<0.01).
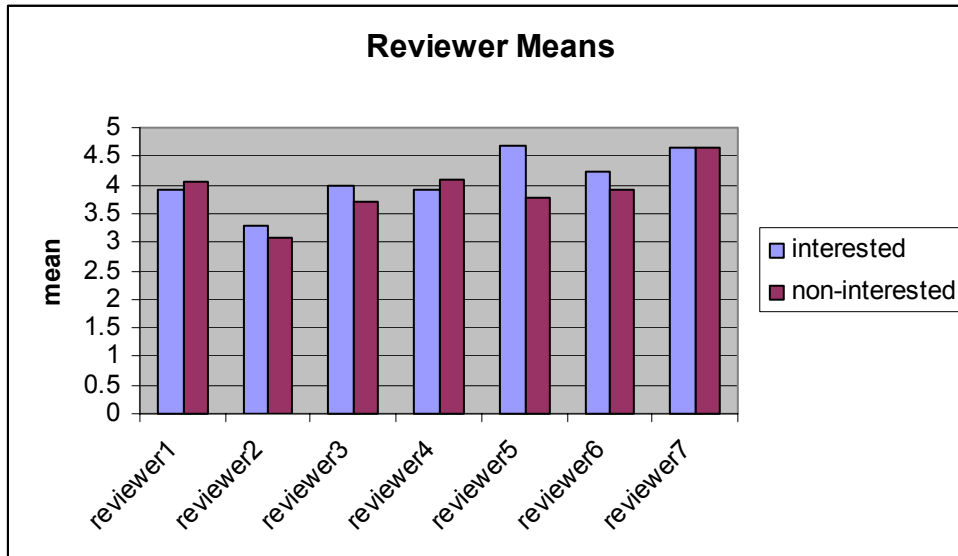
Figure 11:   Reviewer Means

Exploration of quality criteria: Follow up with the experts

We collected the "quality criteria" from 6 out of 7 experts. The following criteria, common for all experts were reported:

- Clear topic

- Up to date information

- Literature cited

- Satisfactory content

- Good lecture flow: ordered, systematic, clearly focused

- The following criteria were reported by just some of the experts

- Personal interest in the topic

- Busy slides (negative)

- Up to date information

- Educational value

- Scientific validity

Thus overall there was little relationship among reviews, in general they had indicated that they were using roughly the same criteria.

Evaluating characteristics of lectures that got highest and lowest scores: descriptive analysis

One of the important aspects of this study was to look at the characteristics of the lectures that were rated the highest and the lowest by the expert reviewers. For all expert reviewers, we looked at 10 lectures that they rated the worst (scores 1-3) and 10 lectures that they rated as the best (score of 5). We also looked at the difference in the scoring pattern of the reviewers from the developing and from the developed countries. The characteristics of these lectures are summarized in the table 10 below.

Table 10: "Good" and "Bad" lecture characteristics

Reviewers	Characteristics of "best" lectures	Characteristics of "worst" lectures
From developing countries N= 3 expert reviewers	Scientific Medical Non-traditional medicine Good flow Simple slides Graphical	Lack of notes Social science Basic methodology oriented Specific to certain area
From Developed countries N=4 expert reviewers	Epidemiologic methods Reputable author Notes Simple slides Simple concepts Good flow Graphical	From developing country Non-traditional medicine Lack of notes Not research oriented Program methodology

The results of this research were most interesting and raised many additional questions that we never thought about before. One may think that graphical lecture with good notes from a reputable author would be rated high by both, reviewer from the developing and developed

country. In reality, even lectures from very reputable authors were oftentimes getting low scores. Lectures judged as "best" by one author, were often rated as "worst" by another. The results of this exercise demonstrate the remarkable inconsistency among reviews and made us ask the question why is there such a high inconsistency and what does it mean for the peer review on the web?

5.   DISCUSSION

The possibility of conducting quality control of research materials entirely online is an attractive feature but relatively unexplored biomedical application of the Internet. Dr. Aaron was one of the first researchers who put a peer review form for the lectures on the web in 1999 and this was one of the first efforts to apply scientific method for the analysis of web based peer review.

The penetrance of the Internet in the population is getting and more and more widespread with scientists world wide almost universally wired (over 70% of the population is connected in the US). This offers the possibility of rapid recruitment of participants and reviewers, and technological advances enable instant collection of data in a secure and confidential manner (Carey, 1997). In this paper we present how web based statistical quality control can become successful and may offer an enticing alternative to peer review systems in the journals. An estimated cost of traditional peer review of an article is $1500, whereas the cost of reviewing Supercourse lecture is basically free (this refers to the general reviews of the Supercourse lectures, not the expert reviews described in this study). Over the course of the past five years, we accumulated several thousand lecture reviews with very little effort. Amazon.com has very similar system of reviews, with similar responses, and it is also free. We are in many ways in the position that manufacturing was 100 years ago. For both, there was no proven quality control

system. Industry forged ahead to find new and better means of QC. Scientific communication has not, until now.

There needs to be hypothesis testing research looking at the quality control systems of the Supercourse project, as well as peer review systems to find out which is better. Research on the value of peer review is limited by the number of factors, including the lack of a validated instrument to measure the quality of reviews (Van Rooyen, 1999)

The current research study was one of the first studies ever conducted that looked at the quality of epidemiological materials on the Internet. One of the unique aspects of this research is that it challenged the traditional paradigm of the peer review system and explored the utilization of an alternative method. Peer review system has been utilized since the times of Aristotle (Barnes, 1981). The *Philosophical Transactions of the Royal Society* is widely accredited as being the first journal to formalize the process of peer review about 300 years ago (Zuckerman, 1971). Despite such a long history in the field of science, recent articles suggest that the whole process of peer review may be in crisis (Mulligan, 2005) and may need to undergo some significant changes. Jefferson's article even suggests that there is very little science behind the peer review process (Jefferson 2002). Our study employed experts who serve as peer reviewers or editors and looked at the process of peer review in lieu of lecture library on the Internet. The advantages of this web based peer review like system are pretty straight forward: eliminating the lag between research and lecture publication time, continuous and evolving quality control process, allowing the author to improve the quality of the lecture, reducing the cost of QC, etc.

The experts who were reviewing lectures for this study came from six countries and from different disciplines within public health: environmental health, cardiovascular health, preventive work, etc. In many ways, expert selection corresponds to the current trends of peer review in the

major biomedical journals: due to difficulties in finding peer reviewers in their own countries, editors often turn for help to reviewers from abroad. Our approach and our expert selection were very sensitive to the needs of the researchers in the developing world. Also, just like in the peer review process, it is impossible to find an expert for lectures that are very specific.

The results of this research were somewhat unexpected, as we thought there would be a high correlation at least among some of the reviewers and a strong relationship to the Supercourse reviews. In retrospect our results were consistent with the existing literature in the area of peer review. A study similar to ours investigated the agreement between two referees when they were evaluating abstracts submitted for a primary care conference based on 4 point scale checklist. The Kappa statistic for inter rater agreement on subjective questions like importance ranged from 0.01 to 0.25, which is similar to the results we received (Montgomery et al, 2002). The agreement among peer reviewers has also been analyzed in the Croatian Medical Journal. Kappa statistic among the peer reviewers was poor to fair for both national and international articles (Marusic et al, 1998).

Outside the field of medical science, we explored the inter rater agreement between Siskel and Ebert, the most popular movie reviewers of the last century in the US. Siskel and Ebert represented the first and most popular of the movie review series genre that emerged on television in the mid-1970s. The lively series focused on the give and take interaction and opinions of its knowledgeable and often contentious co-hosts, Gene Siskel, film critic of the Chicago Tribune and Roger Ebert, film critic of the Chicago Sun-Times. For this dissertation, we looked at the Siskel and Ebert reviews from 1991 to 1996. Interestingly, the agreement between these highly trained reviewers was really low: they disagreed in their ratings in at least 50% of the times.

One may argue that the inter rater agreement in our study was low in our study just because the reviewers were not properly trained to review the materials, however this is unlikely because the expert reviewers that were selected represented a highly experienced group. Moreover, few, if any, reviewers are trained to review articles or grants. The literature in this area suggests that even if you train a reviewer in a group session to do a better job at peer review, there is only a slight impact on the quality of peer review (Schroter et al. 2004); and some studies even suggest that additional training has absolutely no affect on the quality of the review (Callaham 1998, Callaham and Schriger 2002).

The results of this study make us raise a question of whether the Supercourse approach to quality control works and whether or not a similar approach to quality control utilized by Consumer Reports really works. If they work, how exactly and why do they work? The answer to this question is not simple. Consumer Reports have been utilized for quality ratings of various products including healthcare facilities and services. One of the early articles on Consumer Reports in healthcare emphasized the positive aspects of consumer reports on the quality of services. Public release of Consumer Reports may be useful not only in assisting consumers to make informed health care choices, but also in facilitating improvement in the quality of hospital services offered and care provided. (Longo et al, 1997). Later articles in this area sounded much less optimistic. In 2001, Schauffler et al suggested that consumer report cards do not make a difference in decision making, improvement of quality, or competition. The research to date suggests that perhaps we need to rethink the entire endeavor of Consumer Report cards. Consumers desire information that is provider specific and may be more likely to use information on rates of errors and adverse outcomes. Another article published in 2002 suggested that the open access hospital ratings on the Internet poorly discriminated between any 2

individual hospitals' process of care or mortality rates during the study period (Krumholz et al. 2002)

It is very intriguing to see that many people are using Consumer Reports, even though nobody knows if they really work and whether they give an accurate prediction of quality for the products. Consumer Reports probably became so popular because they allow the user to get access to highly customized information: certain products may be of really poor quality in general, however they may possess certain characteristics that may make it valuable to one person in particular. Consumer Reports "work" because they allow the consumer to set their own quality bars and decide what kind of quality parameters they need to look at and evaluate. Usability of Consumer Reports may be a good explanation for their popularity. Just like with the Consumer Reports, it is not clear whether the type of quality control utilized by Amazon.com really works. To our knowledge, there were no formal research studies evaluating quality control in Amazon. Clearly, more research is needed in the area of these popular quality control systems. It was very interesting to see that the mean overall scores of experts was so similar to the mean overall score of non-experts, however the statistics demonstrated that these means are, in fact, different. In this case, it would be interesting to draw a parallel to the field of medicine, where researchers make a distinction between statistically significant difference and clinically meaningful difference. Due to the nature of the research, even under ideal circumstances there is no guarantee that statistical significance has clinical meaning. If the size of a sample is large enough, even trivial differences will be statistically significant. We must remember that statistical significance does not imply that the differences detected between treatment effects are of clinical value. Our case demonstrates a good example of "clinical meaningfulness". Even

though statistically our mean overall scores are different, both of them are very high and demonstrate how positively the Supercourse is viewed by the audience.

In this study, we identified some of the factors that may influence the quality of the peer review mechanism. One of the unique aspects our study looked at was the relationship between the way the expert rates the lecture and his/her personal background. Although our results were not conclusive, several interesting issues came up. Experts may score lectures in his/her area higher than other lectures, just because this expert may better appreciate the content of the lecture. On the other hand, the same expert may be prone to giving lectures in his interest area lower scores, just because he or she may be better equipped to judge the relevance and the novelty of these materials. Future studies definitely need to explore these unique factors influencing individuals' scoring patterns.

The quality of a lecture can be viewed as two separate things: content of the lecture and delivery of the lecture. Interestingly, it was found that factors related to the delivery of the lecture, such as poor grammar, bad spelling or busy slides may have major and dramatic impact on the rating of lecture quality. In our study, reviewers from the developing countries reviewed lectures very differently than the reviewers from the developed countries. The differences were observed not only in the type of lectures that were viewed as "high quality", but also in the digit preference or the way the scores were assigned. Lectures that were written by the authors from the developing countries were oftentimes scored lower, even if the content seemed to be captivating. This does not mean that the researchers from the developing countries are not doing good research. Lack of "traditional" lecture structure for the lectures coming from the developing world and some grammatical mistakes may be some of the reasons why this happened.

We decided to analyze how our findings about the factors that influence peer review process compare to other research studies in the area of peer review. The factors identified in our study, as well as the factors identified by other studies were consolidated in the "peer review factor wheel" presented in the graph. This figure demonstrates the breadth of factors that may play a role in this important process.
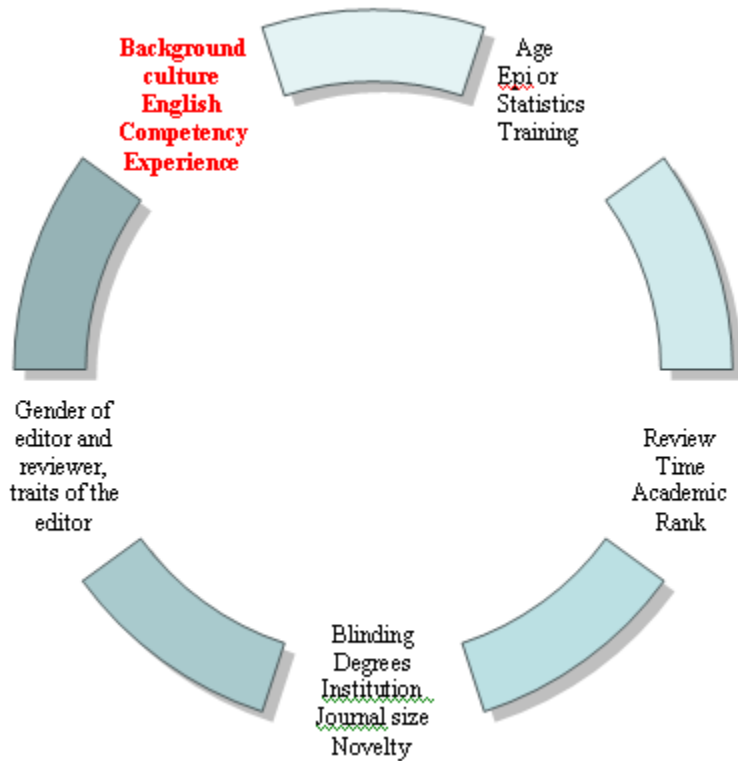
Figure 12:   Peer review factor model

There are several studies evaluating the factors that potentially may play a role in the quality of a peer review. Some of the factors presented in the wheel are cited more often in research articles, such as age and education in the area of epidemiology or biostatistics. Several articles suggest that that younger reviewers tend to give reviews of higher quality (Black et el 1998, Stossel et. Al. 1985) One of the studies suggested that assistant professors or junior faculty

give better reviews (Stossel, 1985). Masking reviewers to author identity as commonly practiced does not improve quality of reviews (Justice et al, 1998). Reviewers with the educational degrees in epidemiology or statistics tend to give better reviews, so now we see in part why as there is so much variability in the system.. Reviewers who are considered to be good come from top academic institutions and are known to the editors (Evans et al, 1993). In general, having more time to conduct the review (up to 3 hours) has also been reported as a factor influencing the quality of the review. Even something like the gender of the editor may potentially be a factor influencing the peer review process (Dickersin et al, 1998). The wheel helps us to make a very important observation: peer review is influenced by so many factors that it is not clear whether any type of inter rater agreement may ever be achieved. Factors highlighted in bold red were the ones that were explored in this study.

It was also interesting to explore the factors that made experts assign lower scores to the lectures. The reason for this difference may arise from the fact that the experts are more content sensitive than the Supercourse reviewers. Supercourse reviewers may rate lectures higher just because they may be fascinated with this wonderful source of information. For many scientists in the developing world, the Supercourse may be the only source of current research information. If somebody like that is serving as a reviewer of the Supercourse lecture, his or her review may be very positive, just because it provides a valuable information resource.

Whereas traditional peer review approaches almost automatically reject any materials coming from the developing world due to poor English or other problems, our project evaluated any materials regardless of the lecture authors' backgrounds. It is interesting to point out that several reviewers critiqued some lectures for poor grammar or style errors. Such comments did not necessarily result in poor overall scores for lectures. What would this mean for the future

developments of peer review and other quality control systems? Is it correct to reject materials from non-native English speakers on the basis of grammar? If a web based journal or Internet based lecture library chooses to accept poorly written material, who should be responsible for editing? These are just some of the questions raised by this research endeavor.

Several factors need to be discussed as possible confounders and addressed in the future investigations. The most important issue that needs to be discussed is the study population. Expert reviewers selected for this research were volunteers. Although the experts had outstanding qualifications and several of them serve as peer review journal editors, all of them have good familiarity with the Supercourse and could have been biased. Thus, the reviews provided by our experts may not necessarily be extrapolated to the "general pool" of the reviewers.

Another issue that needs to be discussed is the fact that all of the reviewers knew that they were participating in the study, thus their reviews might have been influenced by what is called the Hawthorne effect. The Hawthorne effect - an increase in worker productivity produced by the psychological stimulus of being singled out and made to feel important. Individual behaviors may be altered because subjects were aware they were studied. This phenomenon was demonstrated in a research project (1927 - 1932) of the Hawthorne Plant of the Western Electric Company in Cicero, Illinois (Roethlisberger 1939).

Floor and ceiling effects could also be important factors influencing our results. In our case, ceiling effect could result because our lecture scores cannot distinguish between lectures that are somewhat high and those who have very high levels of the construct of quality. Our measure potentially puts an artificially low ceiling on how high a lecture may score and thus could produce bias. Most of the lectures in the Supercourse are relatively good because they are

coming from top academic experts. The scale that we utilized in this research may simply be not sensitive enough to pick out the differences.

Both, floor and ceiling effects are relatively new phenomena discussed in various fields of research. Floor effect is generally defined as the effect of a treatment or combination of treatments that is underestimated because the dependent measure artificially restricts how low scores can be. These interesting factors would need to be addressed in the future research.

One of the interesting findings of this study is the HALO effect that we observed when we looked at the association between the overall score and all other scores for relevance, presentation, etc. The evaluation form we utilized is very short and easy to fill out, however it still takes time to answer 15 questions. Our research demonstrated that it may be possible to replace all of our questions with just one. Simplifying the review process has a major implication for all users of scientific materials on the Internet. Our findings suggest that lectures on the Internet can probably be rated the same way as merchandize in the consumer reports or hotels: good lectures could get a score of five stars and poor lectures can be weeded out by getting a score of one star. Simplifying the peer review process may be the way to go for the editors of the biomedical journals who are unable to attract scientists to review the articles. There is a growing body of literature suggesting that the peer reviewed biomedical journals are experiencing great difficulty finding peer reviewers. Many journals are offering various gifts or even small payments to encourage scientists to review, but the problem remains unsolved. With busy schedules, heavy loads of research work, and constant lack of time, researchers are reluctant to spending too much time reviewing articles. Without simplifying the traditional peer review process it may be impossible for the process to continue. Interestingly, halo effect has been

viewed as a bias and something highly negative in psychological literature and in other areas. In this study, halo may offer an interesting solution to simplifying the lecture review process.

Another issue that needs to be mentioned is the problem associated with lecture selection for our random sample. All foreign language lectures were excluded from this research. The lectures that did not accumulate enough reviews were also excluded. If there was a way to evaluate those excluded lectures, interesting findings could have occured. Purely random sampling is an ideal way to make statistical inferences from the sample, however obtaining a truly random sample is rarely possible in the real practice. What is possible and important is to make sure that the sample selected is not in some way biased (Norusis M.J. 1997) Future quality control studies should place more emphasis on the utility of lectures as a function of quality. It is possible that those lectures that did not get any reviews are of poor quality, so this is something worth exploring in the future.

Why was there such a poor agreement among the reviewers and with the general population of reviewers? Several factors may have played a role in this interesting finding. One of the ways to explain this phenomenon is that it is possible that the individual reviewer's ratings are not consistent over time or have poor intrarater agreement. The reviewer's ratings may differ depending on the scope of expertise of this person. With added skills in certain areas, the ratings of the lectures can go up or down within the same person. Our expert reviewer who evaluated Supercourse lectures in September 2004, may rate the same lectures very differently five years from now. One of the future directions of this research could potentially be in the area of intra rater reliability and figuring out if reviews for the same expert are consistent over time. Another explanation could be that poor inter rater agreement is due to the fact that the expert reviewers all have different occupations. Our previous research demonstrated that medical

doctors tend to give lectures lower scores than professors.   In this case, some of our experts were MDs, some professors, and some public health practitioners. This difference could have caused low agreement.

What is the future of peer review? The process did not undergo too many changes in the past several centuries. The lack of progress in this area is seen by contrasting research 150 years ago with that today. John Snow in 1854 characterized cholera in London; the results were published a year later, and by 1856 the information was in undergraduate curriculums—only two years after the original epidemic (UCLA Department of Epidemiology. John Snow. www.ph.ucla.edu/epi/snow.html (accessed February 21, 2005.) In contrast, research completed in 2005 may not be seen in classrooms for more than five years. It is becoming clear that many scientists feel that the process of peer review needs to undergo some changes. "Many referees feel their reviews would benefit if they had formal training in the review process, received feedback on their reviews, or were able to ask colleagues for opinions on the paper being reviewed. Most reviewers would be willing to sign their reviews and feel that the process should be transparent (Snell 2005). Transparency is something that may help to alleviate some of the biases associated with Consumer Reports and Amazon.com. If the identity of the reviewer is revealed, this may prevent some of the reviewers from abusing the system and submitting multiple positive or negative reviews for a certain product or service.

One of the important points that this research demonstrates is the need for a uniform and easy to utilize method to judge the quality of the data on the internet. In the past decade, a great number of tools claiming to judge the quality of the health related sites emerged on the Internet, with at least 47 of them available in 1998 (Jadad and Gagliardi, 1998), and over 90 in 2002.

Despite their growing numbers, it is not clear whether they are measuring what they claim they measure and whether they are effective.

This was beautifully demonstrated by a study conducted in France. The researchers wanted to develop a simple and easy French Code of Ethics, enabling medical students to judge quality of health information the Internet. After three medical informaticians selected ten criteria from previously established codes of ethics from Europe and the USA, this instrument was tested on a sample of 30 health Internet teaching resources. For the panel of experts, Kappa coefficient for quality rating ranged from $k = -0.19$ and $k = 0.33$, demonstrating poor agreement among the raters (Darmoni et al, 2002). These interesting findings go hand in hand with the findings of our study. Many researchers, organizations, and website developers are exploring alternative ways of helping people to find and use high quality information available on the internet. Whether they are needed or sustainable and whether they make a difference remains to be shown (Gagliardi and Jadad, 2002).

Although this study demonstrated that consumer based evaluation of the epidemiological materials in the context of the Supercourse is possible, it is not completely clear whether this approach could be used for other Internet based libraries. The fact that similar approach is working for consumer products, as demonstrated in Amazon.com, is very promising.

6. CONCLUSIONS

6.1. Public Health Significance

The amount of lectures on the Internet is growing with more and more health professionals, teachers, and educators getting access to materials on the Internet. When a professor in Kenya who does not have access to any biomedical journals hops on the Internet in search of health education information, what would be the quality of the information that he or

she would find? This work was one of the first efforts to analyze the quality of the Internet based PowerPoint lectures. This work was especially important for the field of epidemiology because growing number of epidemiologists worldwide rely on the Internet for the latest research information and teaching resources. The majority of the Supercourse collaborators are epidemiologists and more than half of the lectures in the Supercourse are epidemiological in nature. Growing demand for the epidemiological lectures on the Internet also indicates that there is a need for better quality control.

We concluded that just like in the traditional peer reviewed journal, the inter rater agreement among expert reviewers of the epidemiological lectures is not high. We also concluded that one score may be sufficient for the lecture evaluation vs. five or more scores. Overall, our study demonstrated that Supercourse is viewed very positively by both, lay audience and well established expert users. The study also helped to identify several interesting factors that may influence the peer review process, such as reviewer's background.

In the past few years, there has been a push to provide free health information on the Internet through open source free web based journals (Eysenbach 2004). A journal like Journal of Medical Internet Research is free of charge and available on the Internet to anybody with the modem and a web browser. Traditional peer review mechanisms are still utilized in these journals and authors need to pay if their article is accepted for publication. This may prevent the researchers from the developing world from publishing their data. Additionally, even though peer reviewed processes are "expedited" in open access journals, they still take quite a while (about 4 weeks). Can we still use traditional peer review mechanisms to judge the quality of the lectures on the Internet? Probably not, considering the fact that it has many biases, consumes too much time, and there is no conclusive scientific date on its effectiveness. Reputable open access

journals represent only a miniscule fraction of the total amount of health related information on the Internet. In the field of public health and medicine, we need a way to access the quality of all information on the internet because of its crucial importance to the health of people.

In this study, we were on the journey to new directions of quality control: quality control driven by expert consumers of such information. The importance of this study is evident when you think about the number of people turning to the Internet for health related information and the number of instructors turning to the same source for teaching materials that will enhance their curricula.

6.2. Future directions

The findings of this research were quite intriguing, but it was just a small part in a big puzzle in the fate of the peer review system and its future development. Obviously, the system of peer review is not functioning in the way it is expected to function: "peers" do not agree in their quality judgments not only in the journal articles, but also in the Internet lectures. There are 2 possible ways in which peer review research can develop in the future:

- More studies can be conducted on the validity of peer review, looking at more aspects of the system and identifying more strengths and weaknesses

- Accepting the fact that traditional peer review may not work and move it to a next level or to a brand new system

Doing additional studies on peer review may be a good way to go, but it seems like it may not lead to any useful findings. Many research groups nationally and internationally, and completely independently of each other, found out that there is generally poor agreement among peer reviewers in the paper journals. Having additional studies done in the area of paper based journal peer review seems to be a waste of valuable resources. Since there is lack of studies on

peer reviewer agreement for Internet based materials, it would probably be useful to carry out additional study to confirm the results of this research. The nature of Internet would allow to carry out a study looking at multiple peers in multiple settings. If this route of research is taken, it may be interesting to conduct further research to compare the review trends of the Supercourse lectures with the review trends of the popular system of Amazon.com. Preliminary observations our group has conducted in this area suggested that these patterns may be very similar. For example, we found that both in Amazon.com and in the Supercourse not all lectures and books available in the library are rated. It would be interesting to find out more about the characteristics of books and lectures that never undergo peer review.

However, instead of refining the research that was already done, it would be more interesting to investigate alternative means to the peer review mechanisms completely separate from existing process. There is a very limited body of literature available on this topic, as very few people have dared to challenge the framework that has been in place for hundreds of years. One of the authors suggested that refereed journal literature needs to be freed from both paper and its costs, but not from peer review, whose "invisible hand" is what maintains its quality (Harnad 2000). This author suggested that peer review should remain basically the same, but move to the Internet media. With the ever growing nature of the Internet, this approach may not be feasible.

There were also more radical suggestions, exploring the brand new alternatives to the peer review mechanisms. One of the most radical ideas is to let every submission be published and let the reader decide what is to be taken seriously. This would amount to discarding the current hierarchical filter -- both its active influence, in directing revision, and its ranking of quality and reliability to guide the reader trying to navigate the ever-swelling literature

(Hitchcock et al. 2000). By examining literature in this area and by looking at the finding of this research study, we would like to argue in favor of this radical suggestion. Instead of trying to find ways to prevent people from publishing their data, we need to work on better retrieving mechanisms that would help the reader to sort through the information available in the electronic journals and on the Internet in general.

Although radical, this approach would allow the researchers from the developing world publish their data, without the need to struggle with peer review mechanisms. Currently, researchers from the developing world are basically excluded from publishing in the leading biomedical journals and their valuable research information is rarely shared with their western peers. Partly because of the peer review mechanisms, science is dominated by a few countries and the contribution of the researchers from the developing world is simply forgotten (Gibbs 1995). A similar situation is observed in the peer review of grant proposals at NIH and other agencies.

Another feasible approach to quality control of the materials on the Internet is the creation of a Google like system, where items that utilized the most (measured by hits to the website) and those that have the highest number of sites linking to them are labeled as high quality materials. Just like any other system, this system is not perfect and may be abused. However, it is clear that Google system may be good for rapid quality control for large numbers of materials. With the growing and ever evolving nature of the Internet, Google-like quality control system may be the answer to many of the existing problems related to the quality and usability of the Internet based materials in the Supercourse and other web libraries.

The optimal quality control system of epidemiological materials on the Internet would probably combine several aspects of the existing quality control system utilized in the

biomedical journals and other areas. For example, Multimedia Educational Resource for Learning and Online Education (MERLOT) (http://www.merlot.org) combines the system of traditional peer review and consumer feedback, similar to that in amazon.com. MERLOT is a free and open resource designed primarily for faculty and students of higher education. Links to online learning materials are collected at MERLOT website along with annotations such as peer reviews and assignments.

Change in the process of peer review will not be possible without challenging traditional paradigms and exploring new alternatives. Recent article in the New England Journal of Medicine suggested that the medical libraries the way we know them will not survive for long (Lindberg and Humhreys, 2005). The same may be true for peer review and quality control. This research made it clear that the scientific community, especially in the area of medicine, is in need of an improved science of quality control. We need a better definition of quality and better way to implement the quality control mechanisms. Exploration of enhanced information retrieval mechanisms for the Internet based lectures and articles could be one of the first steps towards a better alternative in quality control.

Implementation of new quality control mechanisms for biomedical literature and web materials will need to engage all the stakeholders involved in this process. Quality control mechanisms have been successfully implemented in the industry because everybody: consumers, companies, and workers demanded high quality products and safe work environments. As W.E. Deming points out (written communication, November 1987), even 99.9% success rate in the industry may not be good enough: "If we had to live with 99.9%, we would have: 2 unsafe plane landings per day at O'Hare, 16,000 pieces of lost mail every hour, 32000 bank checks deducted from the wrong bank account every hour" (Leape 1994). Medical workers, public health

professionals, journal editors, etc. may be somewhat reluctant to accept the changes to the peer review system because they do not like to admit the fact that they can err. This situation can change with the development of the science of quality control for scientific publications on paper and on the web. Our ultimate goal should be to bring the success of quality control in the industry to biomedical journals.

APPENDIX A

Complete review form utilized for data collection

Peer Review of the Lecture

Your input is critical to the continued development of the Supercourse. Please complete the review form below and return your response by clicking the **submit button** at the bottom of this page.

1. Name:

2. Position:

3. Organization:

4. Email:

5. Have you ever taught an Introductory Course in Epidemiology?

   Yes

   No

6. Do you currently teach an Introductory Course in Epidemiology?

   Yes

   No

7. How interested would students be in this lecture?

   Very

   Some   What

   Little

   Not At All

8. May we post your review on this web site?

   Yes

   No

9. Did the graphics transfer in a reasonable amount of time?

☐ Yes

☐ No

Please rate the lectures on the following characteristics:

5 = Excellent,	4 = Above Average,	3 = Average,	2 = Below Average,	1 = Poor

10. Content: ☐ 5 ☐ 4 ☐ 3 ☐ 2 ☐ 1

11. Presentation: ☐ 5 ☐ 4 ☐ 3 ☐ 2 ☐ 1

12. Relevance: ☐ 5 ☐ 4 ☐ 3 ☐ 2 ☐ 1

13. Overall Rating: ☐ 5 ☐ 4 ☐ 3 ☐ 2 ☐ 1

---

14. How does the quality of the lecture compare with your expectations about it?

☐ (5) Well above what I expected

☐ (4) Above what I expected
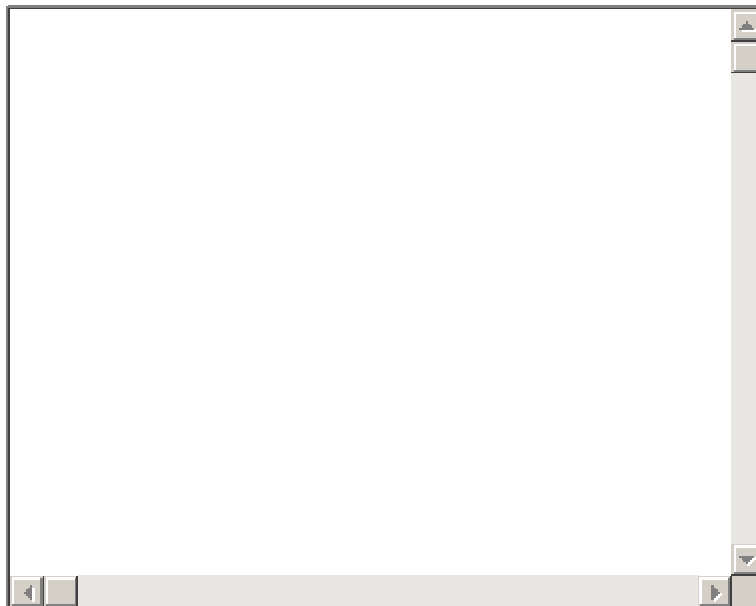
☐ (3) Same as expected

☐ (2) Somewhat below what I expected

☐ (1) Well below what I expected

---

15. Please provide your general and specific comments about the lecture. You amy provide web sites you konw which are appropriate to the lecture below:

Submit button

If you need to change your ratings or comments, please use the reset button:

Reset button

APPENDIX B
List of lectures that were evaluated by the expert reviewers

1. World Wide Web (WWW) as a Global Virtual Library
2. The Newcastle Critical Appraisal Worksheet. A format for examining journal articles.
3. Income Inequality and Mortality in Canada and the United States
4. Cryptosporidium: The Milwaukee Case
5. Epidemiology and Management of Diarrheal Diseases
6. Terrorism: the Epidemiology of Fear. Part I
7. Investigating an Outbreak. Part I
8. Principles of Public Health - The Mission, Core Functions and Ten Essential Services. Part I.
9. Recent Trends in Diet and Serum Lipids Level in Japan
10. Domestic Violence
11. Disability Adjusted Life Years Possibilities and Problems
12. Virus Replication (Veterinary Virology).
13. An Approach to Management. The Special Problems of Health Services
14. Teenage Driver Crashes -- Carrying Passengers as a Risk Factor
15. Introduction to Focus Groups
16. The Herbal Treatment of Diseases
17. Health Promotion (an overview)
18. Self-Rated Health in Epidemiological Surveys as a Predictor of Disability and Mortality
19. A Brief Introduction to Epidemiology - V (Principles of Organizing & Presenting Epidemiologic Data)
20. Biomechanical Considerations for Rehabilitation of the Knee
21. Introduction to Health Economics
22. Disease Categorization and Classification
23. Living and dying: Health, Illness and Disease
24. Health, Nutrition in Islam
25. Epidemiology of Endemic Fluorosis
26. Terrorism: the Epidemiology of Fear. Part II
27. A smoking gun? Detecting causes of disease
28. Toxicology and Risk Assessment. (3rd of 10 Lectures on Toxicological Epidemiology)
29. Epidemiology of Diabetes Complications - I
30. Case-Control Studies
31. Public Health in Cuba
32. A Primer on Sample Survey
33. Epidemiologic Transition
34. Investigating an Outbreak. Part II
35. Herd Immunity and Vaccination
36. Population Sampling
37. Case Crossover Design
38. The Big Ten Tropical Diseases. Categorization and research strategic emphases
39. Cancer Trends in England and Wales
40. Evidence based health care (EBHC)
41. Natural History and Determinants of Type 2 Diabetes

BIBLIOGRAPHY

Anastasi A. Psychological Testing. New York: McMillan, 1982

Ary D, Jacobs LC, Razavieh A. Introduction to Research in Education. New York, NY: Hot, Rinehart and Winston, 1985

Assael H. Consumer Behavior and Marketing Action. Cinncinati, Ohio: South-Western College Publishing, 1995

Nua Internet Surveys http://www.nua.ie/surveys/how_many_online/

Aaron DJ, Sekikawa A, Acosta B, Sa ER, LaPorte RE. Transnational education: The Global Health Network Supercourse (www.pitt.edu/~super1) {Procedeengs}. XX World Congress of Pathology and Laboratory Medicine, Sao Paulo (Brazil), 1999, p 129-133

Acosta B. Breaking the Language Barrier. British Medical Journal http://www.bmj.com/cgi/eletters/313/7067/1264/b#EL1 (electronic response, June 1999)

Barkman WE. In-Process Quality Control for Manufacturing. New York: Marcel Dekker Inc., 1989.

Barnes J. Proof and the Syllogism. in: Berti. Discusses the principles of Aristotle's endoxos 1981:17–59.

Beckwith NE, Kassarjian HH, Lehmann DR. Halo effects in marketing research: review and prognosis. Advances in Consumer Research 1978: 465-7.

Berland GK. Health information on the internet: accessibility, quality, and readability in English and Spanish. JAMA 2001; 285: 2612-2621

Beecher DE. The evaluation of teaching. Syracuse University Press, 1994

Besterfield Quality Control, sixth edition. New Jersey: Upper Saddle River, 2001

Cavalli P. False-negative results in Down's syndrome screening. Lancet 347: 965-966, 1996

Black N, Van Rooyen S, Godlee F, Smith R, Evans S. What makes a good reviewer and a good review for a general medical journal? JAMA Jul 15;280(3):231-3, 1998

Cavalli P. False-negative results in Down's syndrome screening. Lancet 347: 965-966, 1996

Carey VJ. Using hypertext and the Internet for structure and management of observational studies. Stat Med 16: 1667-82, 1997

Cross SS Kappa statistics as indicators of quality assurance in histopathology and cytopathology J Clin Pathol.Jul;49(7):597-9, 1996

Callaham ML, Wears RL, Waeckerle JF. Effect of attendance at a training session on peer reviewer quality and performance. Ann Emerg Med. 32(3 Pt 1):318-22, 1998

Callaham ML, Schriger DL. Effect of structured workshop training on subsequent performance of journal peer reviewers. Ann Emerg Med. 40(3):323-8, 2002

Darmoni SJ, Le Duff F, Joubert M, Le Beux P, Fieschi M, Weber J, Benichou J A preliminary study to assess a French code of ethics for health teaching resources on the Internet. Stud Health Technol Inform. 90:621-6, 2002

Delamothe T, Twenty Thousand Conversations BMJ 324:1171-1172, 2002

Dhillon BS. Quality Control, Reliability, and Engineering Design. New York: Marcel Dekker Inc., 1985

Dickersin K, Fredman L, Flegal KM, Scott JD, Crawley B. Is there a sex bias in choosing editors? Epidemiology journals as an example. JAMA. 280(3):260-4, 1998

Eaton L. A third of Europeans and almost half of Americans use internet for health information BMJ 325: 989, 2002

Evans AT, McNutt RA, Fletcher SW, Fletcher RH. The characteristics of peer reviewers who produce good-quality reviews. J Gen Intern Med. 8(8):422-8, 1993

Eysenbach G. Peer-review and publication of research protocols and proposals: a role for open access journals. J Med Internet Res. 2004 Sep 30;6(3):e37.

Forsström J. Why certification of medical software would be useful? Int J Med Inf 1997; 47: 143-152

Gibbs WW. Lost science in the Third World. *Scientific American*, p.92-99, August 1995

Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. BMJ. 2002 Mar 9;324(7337):569-73.

Harmon A. Amazon Glitch Unmasks War of Reviewers The New York Times February 14, 2004

Harnad S. The Invisible Hand of Peer Review, *Exploit Interactive*, issue 5, April 2000

Hilsenbeck SG, Glaefke GS, Feigel P, Lane WW, Golenzer H, Ames C, Dickson C. Quality Control for Cancer Registry. Washington, D.C.: U.S. Department of Health and Human Services, 1985

Hitchcock S, Carr L, Jiao Z, Bergmark D, Hall W, Lagoze C, Harnad S. (2000) Developing services for open e-print archives: globalization, integration and the impact of links.

Proceedings of the 5th ACM Conference on Digital Libraries. San Antonio Texas June 2000.

http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.acm.htm

Hoffman-Goetz L, Clarke JN. Quality of breast cancer sites on the World Wide Web. Canadian Journal of Public Health. 91(4):281-4, 2000

Impiccatore P, Pandolfini C, Casella N, Bonati M. Reliability of health information for the public on the world wide web: systemic survey of advice on managing fever in children at home. BMJ 314: 1875-1879, 1997

Jadad AR, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? JAMA. 279(8):611-4, 1998

Jefferson T, Alderson P, Wager E, Davidoff F. Effects of editorial peer review: a systematic review. JAMA. 2002 Jun 5;287(21):2784-6.

Justice AC, Cho MK, Winker MA, Berlin JA, Rennie D. Does masking author identity improve peer review quality? A randomized controlled trial. PEER Investigators. JAMA 1998 Sep 16;280(11):968.

Kim S, Lemeshow S, Difficulties of using kappa statistics in epidemiologic studies 129th meeting of APHA Abstract #26701, 2001

Kirkpatrick DL. Evaluating Training Programs: The Four Levels. San Francisco, CA: Berrett-Koehler, 1994

Kowalski JP. Evaluating Teacher Performance. Arlington, Va: Educational Research Service, 1978

Krumholz HM, Rathore SS, Chen J, Wang Y, Radford MJ. Evaluation of a consumer-oriented internet health care report card: the risk of quality ratings based on mortality data. JAMA. 2002 Mar 13;287(10):1277-87.

Leape LL Error in Medicine Journal of the American Medical Association 272(23): 1851-57

Leuthesser L, Kohli C, Harich K, Brand equity: the halo effect measure European Journal of Marketing, May 1995, vol. 29, no. 4, pp. 57-66(10)

Lindberg D, Humphreys B. 2015—The future of medical libraries. New England Journal of Medicine 352;11: 1067-1070

Longo DR, Land G, Schramm W, Fraas J, Hoskins B, Howell V. Consumer reports in health care. Do they make a difference in patient care? JAMA. 1997 Nov 19;278(19):1579-84.

Lundberg G The "omnipotent" Science Citation Index Impact Factor MJA 2003 178 (6): 253-254

Meadow R. A case of murder and the BMJ. BMJ 2002; 324: 41-43

Morrison J. ABC of learning and teaching in medicine: Evaluation
BMJ, Feb 2003; 326: 385 – 387

Montgomery AA, Graham A, Evans PH, Fahey T. Inter-rater agreement in the scoring of
abstracts submitted to a primary care conference BMC Health Serv Res. 2002 Mar
26;2(1):8.

Marusic A, Mestrovic T, Petrovecki M, Marusic M. Peer review in the Croatian Medical Journal
from 1992 to 1996. Croat Med J. 1998 Mar;39(1):3-9.

Morrison J. ABC of learning and teaching in medicine: Evaluation
BMJ, Feb 2003; 326: 385 – 387

Mulligan A. Is Peer Review in Crisis? Oral Oncology 2005 Feb;41(2):135-41.

Norusis MJ SPSS 7.5 Guide to Data Analysis. A Simon&Schuster Company, New Jersey, 1997

Oermann MH, Lesley M, Kuefler SF. Using the Internet to teach consumers about quality care.
Jt Comm J Qual Improv. 2002 Feb;28(2):83-9

Oliver M. and Conole G. Evaluating Communication and Information Technologies: A Toolkit
for Practitioners. Active Learning 8, Institute of Learning and Teaching, 1998

Rigby M, Forsstrom J, Roberts R, Wyatt J. Verifying quality and safety in health informatics
services. BMJ 2001; 323: 552-556

Roethlisberger FJ, Dickson WJ. Management and the Worker Boston, Mass.: Harvard University
Press, 1939

Rogers R. A global information society for health▬recommendations for international action. Br
J Healthcare Computing Information Manage 1999; 16: 28-30

Sekikawa A, Aaron DJ, Acosta B, Sa ER, LaPorte RE. Does the perception of web page
downloading speed influence the evaluation of health content? Public Health 2000

Seidman JJ, Steinwachs D, Rubin HR. Conceptual framework for a new tool for evaluating the
quality of diabetes consumer-information Web sites. J Med Internet Res 2003 Oct-
Dec;5(4):e29

Snell L, Spencer J. Reviewers' perceptions of the peer review process for a medical education
journal. Med Educ. 2005 Jan;39(1):90-7.

Smith R, Milton and Galileo would back BMJ on free speech Nature 427: 287, 2004.

Stossel TP. Reviewer status and review quality: experience of the Journal of Clinical
Investigation. N Engl J Med. 1985;312:658-659.

Schroter S, Black N, Evans S, Carpenter J, Godlee F, Smith R. Effects of training on quality of peer review: randomised controlled trial. BMJ. 2004 Mar 20;328(7441):657-8.

Thorndike, E. L. A constant error in psychological ratings. Journal of Applied Psychology, 4, 469-477, 1920

Van Rooyen S, Black N, Godlee F Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. J Clin Epidemiol. 1999 Jul;52(7):625-9.

Weisbord SD, Soule JB, Kimmel PL. Poison on line—acute renal failure caused by oil of wormwood purchased through the internet. N Engl J Med 1997; 337: 825-827

Westgard JO Six Sigma Quality Design and Control 2001

Wyatt JC. Measuring quality and impact of the World Wide Web [commentary]. BMJ 1997; 314: 1879-1881

Wilkinson P. Down's test leaves 150 women in abortion fear. Times, 2000 May 31: 1, 3.

Ziebland S, Chapple A, Dumelow C, Evans J, Prinjha S, and Rozmovits L
    How the internet affects patients' experience of cancer: a qualitative study
    BMJ, Mar 2004; 328: 564 - 0.

Zuckerman H, Merton RK, Patterns of evaluation in science: Institutionalization, structure and functions of the referee system, Minerva 9 (1971) (1), pp. 66–100.