University of Rhode Island
DigitalCommons@URI

Open Access Dissertations

2017

Assessing Reading Grade Level of Online Mental Health Materials: Practical and Methodological Considerations

Dorothy D. Skierkowski
University of Rhode Island, dskierkowski@uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Terms of Use
All rights reserved under copyright.

Recommended Citation
Skierkowski, Dorothy D., "Assessing Reading Grade Level of Online Mental Health Materials: Practical and Methodological Considerations" (2017). *Open Access Dissertations.* Paper 580.
https://digitalcommons.uri.edu/oa_diss/580

This Dissertation is brought to you by the University of Rhode Island. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu. For permission to reuse copyrighted content, contact the author directly.

ASSESSING READING GRADE LEVEL OF ONLINE MENTAL HEALTH MATERIALS: PRACTICAL AND METHODOLOGICAL CONSIDERATIONS

BY

DOROTHY D. SKIERKOWSKI

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2016

DOCTOR OF PHILOSOPHY DISSERTATION

OF

DOROTHY D. SKIERKOWSKI

APPROVED:

Dissertation Committee:

Major Professor Paul Florin

Lisa Harlow

Yinjiao Ye

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2016

ABSTRACT

The Internet can be conceptualized as a useful tool for providing people with a vast array of mental health information at the click of a button. Despite this plethora of available knowledge, oftentimes the information that is presented on popular physical and mental health websites is written for an audience with a reading grade level higher than the national 6^{th}-8^{th} grade average. Although the CDC has developed guidelines for developing online patient health materials that account for disparities in health literacy across various socio-demographic groups, adherence to these guidelines is largely poor and minimally monitored. This discrepancy can have broad public health implications when considering the suggested relationship between low health literacy and poor health outcomes.

The present study systematically examines grade level readability scores for online information describing sixteen different mental health disorders, extracted from six highly utilized mental health websites, using a general estimating equations approach. In order to best understand this problem, two manuscripts are presented herein. The first manuscript focuses on public health concerns associated with higher than average reading grade level estimates of online mental health materials, whereas the second manuscript focuses on the methodology used to make these determinations. Results suggest that reading grade level estimates of publicly available online mental health information are much higher than the 6^{th} – 8^{th} grade levels suggested by the CDC, such that the average reader will not be able to effectively understand the selected text. This finding can have broad implications from a public health perspective and maintain existing health disparities.

ACKNOWLEDGMENTS

First and foremost, the author would like to acknowledge Dr. Paul Florin for his inspiration and assistance with writing this text. Likewise, many thanks are extended to Dr. Jason Machan and Dr. Lisa Harlow for their invaluable mentorship and guidance throughout the dissertation writing process. The author would also like to acknowledge the assistance of Dr. Yinjiao Ye, Dr. Minsuk Shim, and Dr. Leslie Mahler for their contribution to this final product.

PREFACE

The manuscript format is in use. Two manuscripts are presented herein. The first manuscript describes in detail the problem under investigation and the results of the study conducted by the authors, and will be submitted to the American Journal of Public Health. The second manuscript describes the methodology used in the study in greater detail, with a particular emphasis on how these methods can be applied by public health and/or psychology researchers. The second manuscript will be submitted to the journal of Evaluation and the Health Professions.

TABLE OF CONTENTS

LIST OF TABLES/FIGURES

TABLES PAGE

FIGURES

MANUSCRIPT- 1

To be submitted to The American Journal of Public Health

Assessing grade level readability of online mental health materials using a Generalized Estimating Equations approach: Implications for public health and beyond

Dorothy D. Skierkowski, MA[1], Paul Florin, PhD[1], Lisa Harlow, PhD[1], Jason Machan, PhD[2], Minsuk Shim, PhD[3], Yinjiao Ye, PhD[4], Leslie Mahler, PhD[5]

[1]Department of Psychology, University of Rhode Island, Kingston, RI, USA

[2]Director of Biostatistics Core, Lifespan Corporation, Rhode Island Hospital, Providence, RI, USA

[3]Department of Education, University of Rhode Island, Kingston, RI, USA

[4]Department of Communication Studies, University of Rhode Island, Kingston, RI

[5]Department of Communication Disorders, University of Rhode Island, Kingston, RI

Corresponding Author: Dorothy D. Skierkowski, MA

Department of Psychology

Community Research and Services Team

University of Rhode Island

80 Washington Street

Providence, RI, 02903

Phone: +1-860-692-8209

Email: dskierkowski@uri.edu

Abstract

The Internet can be conceptualized as a useful tool for providing people with a vast array of mental health information at the click of a button. Despite this plethora of available knowledge, oftentimes the information that is presented on popular physical and mental health websites is written for an audience with a reading grade level higher than the national 6^{th}-8^{th} grade average. Although the CDC has developed guidelines for developing online patient health materials that account for disparities in health literacy across various socio-demographic groups, adherence to these guidelines is largely poor and minimally monitored. This discrepancy can have broad public health implications when considering the suggested relationship between low health literacy and poor health outcomes. The present study systematically examined grade level readability scores for online information describing sixteen different mental health disorders, extracted from six highly utilized mental health websites, using a general estimating equations approach.

Introduction

In light of the massive expansion of the Internet over the last decade, a plethora of information has now become available on almost any topic imaginable. Given the existing socioeconomically and geographically- based health disparities in the United States, this treasure trove of knowledge can help to inform decision-making on important physical and mental health topics ranging from signs and symptoms of heart disease, to mental health concerns such as substance use and anxiety. Indeed, in a few short keystrokes, people now have access to a myriad of information from multiple sources via popular online search engines such as Google, Bing, Yahoo, and Ask.com.

However, despite the popularity of online health materials as a vital source of information from which to make important health-related decisions, little attention has been paid to the *readability* of these materials, where readability refers to a systematic measure of ease with which a passage of text can be read (Albright, de Guzman, Acebo, Paiva, Faulkner, & Swanson, 1996; McInnes & Haglund, 2011). This lack of attention to readability of online public health information is particularly problematic considering that approximately 35% of US citizens have basic and below basic health literacy, 53% have intermediate health literacy, and only 12% have proficient health literacy. Here, health literacy is defined as the ability to search for, comprehend, and utilize written health education materials to make educated healthcare decisions (Berkman, Sheridan, Donahue, Halpern & Crotty, 2011; Berkman, Sheridan, Donahue, Halpern, Viera, & Crotty, et al., 2011; Kutner, Greenberg, Jin & Paulsen, 2006; Committee on Health Literacy, 2004).

The purpose of this report is to compare the current readability estimates of several popular mental health-related topics from various online sites, to determine whether there are systematic differences in grade-level readability based on topic and/or source from which the information was obtained. The following disorders were considered for analysis based on their 12-month prevalence rates for adults within the United States population: specific phobia – 8.7% (National Institute of Mental Health, 2016); substance abuse/addiction – 8.2% (Substance Abuse and Mental Health Services Administration, 2014); alcohol abuse/alcoholism – 6.8% (National Institute on Alcohol Abuse and Alcoholism, 2016); social phobia – 6.8% (National Institute of Mental Health, 2016); major depressive disorder (MDD) – 6.7% (Center for Behavioral Health Statistics and Quality, 2015); attention deficit/hyperactivity disorder (ADHD) – 4.1%; post-traumatic stress disorder (PTSD) – 3.5%; generalized anxiety disorder (GAD) – 3.1%; panic disorder – 2.7%; bipolar disorder – 2.6%; borderline personality disorder – 1.6%; schizophrenia – 1.1%; obsessive compulsive disorder (OCD) – 1.0%; agoraphobia - .8%; bulimia nervosa - .3%; and anorexia nervosa – lifetime prevalence .6% (National Institute of Mental Health, 2016).

Results from this study can be used to create a general set of readability guidelines from which to modify existing online mental health-related materials and/or compile new information in a manner that is consistent with the reading level and education of the general population. Indeed, it is possible that by illustrating the educational bias inherent in much of the written mental health information that is available online, we can begin to address ways in which to reduce this gap and serve those who are most in need (World Health Organization, 2010).

Although there is some research examining readability scores for a range of physical conditions (see, for example, Brigo, Otte, Igwe, Tezzon & Nardone, 2015; Colaco, Svider, Agarwal, Eloy & Jackson, 2013; Misra, Agarwal, Kasabwala, Hansberry, Setzen & Eloy, 2013; or Svider, Agarwal, Choudhry, Hajart, Baredes, & Liu et al., 2013), little if no attention has been paid to assessing readability of online mental health materials. Furthermore, only one study to date has explored the topic of readability using a mixed modeling approach (McEnteggart, Naeem, Skierkowski, Baird, Ahn & Soares, 2015). Hence, this study is novel in that it is the first of its kind to explore the readability of mental health-related information for 16 of the most prevalent mental health disorders, using data extracted from several of the most popular mental health websites, using multiple readability indices.

Who uses the Internet in the United States?

According to a recent study by the Pew Research Center (Perrin & Duggan, 2015), 84% of all Americans use the Internet. Given the heterogeneity of the U.S. population, as well as vast differences in access to technological resources across various socioeconomic spheres, it is important to further examine rates of use by level of education, income, race/ethnicity, gender, and age. For instance, 95% of college-educated Americans use the Internet, as compared with 90% of those with some college education, 76% of those with a high school degree, and 66% with less than a high school diploma (Perrin & Duggan, 2015).

Likewise, 95 - 97% of those earning more than $50,000 per year are Internet users, as compared with 85% of individuals earning between $30,000 - $49,999, and 74% earning less than $30,000 annually. Despite these gaps, there has been much

growth in Internet use over the past 15 years among those in lower-income households and lower levels of educational attainment, such that class differences have shrunk somewhat and many are now able to regularly access this resource from a range of technological platforms (Perrin & Duggan, 2015).

Examination of Internet use by race/ethnicity reveals that 97% of English-speaking Asian individuals use the Internet regularly, as compared with 85% of non-Hispanic Whites, 81% of Hispanics, and 78% of non-Hispanic Blacks. Similar rates of use are evidenced across genders, with 85% of men, and 84% of women indicating Internet use (Perrin & Duggan, 2015). Lastly, a breakdown of use by age indicates that 96% of adults ages 18-29, 93% of adults ages 30 – 49, 81% of adults ages 50-64, and 58% of adults ages 65 or older are Internet users. Although older adults have traditionally been the slowest age group to adopt this technology, a majority of senior citizens now indicate regular Internet use (Perrin & Duggan, 2015).

Despite some differences in rates of adoption among these heterogeneous groups, it is fair to state that a majority of Americans are using the Internet on a regular basis. Hence, there is much potential to utilize this tool to empower people to make more informed choices about their mental health care needs. However, in order to make specific recommendations and develop an action plan for increasing access to, and comprehension of, online mental health materials, it is first important to examine *how* people are currently seeking health information on the Internet, as well as how these behaviors are related to users' general sense of health literacy.

How are people using the Internet to acquire health-related information?

According to a 2012 study conducted by the Pew Research Center, approximately 72% of Internet users reported seeking health information online within the past 12 months (Fox & Duggan, 2013). Likewise, 77% of online health seekers reported beginning their search at a general search engine website such as Google, Bing, or Yahoo, whereas approximately 13% reported beginning at a more specialized medical website such as WebMd.com. Furthermore, 55% of users reported searching for a specific disease or medical problem, and 43% reported searching for a certain medical treatment or procedure. Approximately half of users reported searching for a close family member or friend (Fox & Duggan, 2013).

In addition, 35% of U.S. adults indicated that they have specifically gone online to find out what condition they or someone else might have, and 46% of these 'online diagnosers' reported that the information obtained led them to think they needed medical intervention (Fox & Duggan, 2013). The remaining 38% reported saying that they could take care of the issue themselves at home, with 11% being ambivalent about the decision to seek additional medical care. Participants also reported on the accuracy of their initial diagnosis, with 43% indicating that a medical professional confirmed or partially confirmed their hypothesis, 35% indicating they did not visit a professional, and 18% indicating that a medical professional either disagreed with the initial diagnosis or offered an alternate medical opinion (Fox & Duggan, 2013).

Health literacy

These statistics are important when considering the potential gravity of misdiagnosing or ignoring a serious medical problem based on written information

obtained online, particularly when this information is only readable by a small fragment of the population. Indeed, given that approximately 77 million Americans have basic to below basic health literacy, defined as the ability to read, understand, locate, and interpret health-related information correctly in text (America's Health Literacy, 2008), and that the average reading level across the United States is no higher than the 6^{th} - 8^{th} grade (Kutner, Greenberg, Jin, & Paulsen, 2006; Paasche-Orlow, Parker, Gazmararian, Nielsen-Bohlman & Rudd, 2005), it is important that health information be written at a level that is accessible by the majority of consumers.

According to the U.S. Department of Health and Human Services' Office of Disease Prevention and Health Promotion report on Health Communication Activities (2008), results from the National Assessment of Adult Literacy survey suggest that health literacy is an issue for all racial and ethnic groups, with 28% of Whites, 57% of Blacks, 65% of Hispanics, and 34% of Others (including Asians, Native Americans, and multi-racial adults) in the basic to below basic health literacy groups. Within the scope of this study, health literacy was defined as the ability to successfully: read a set of short instructions and identify what is permissible to drink before a medical test (below basic health literacy); read a pamphlet and give two reasons why a person with no symptoms should be tested for a disease (basic health literacy); read instructions on a prescription label and determine at what time a person can take the medication (intermediate health literacy); and, using a table, calculate an employee's share of health insurance costs for one year (proficient health literacy).

Results from this study also indicated that lower health literacy is associated with less education: 76% of individuals with less than a high school degree, 44% of

those with a high school diploma, 21% of those who had completed some education beyond high school, and 12% of those with a Bachelor's degree or higher, were at the below basic or basic levels for health literacy. Likewise, uninsured adults (53%) and those enrolled in Medicare (57%) and Medicaid (60%) were more likely to be at the below basic or basic levels than those who received insurance from an employer (24%). Interestingly, only 15% of adults with below basic health literacy indicated using the Internet "some" or "a lot" of the time for obtaining health information, as compared with 31% of those with basic health literacy, 49% with intermediate health literacy, and 62% of those with proficient health literacy (America's Health Literacy, 2008). Clearly, marketing online health information for the 12% of users who possess proficient health literacy only serves to perpetuate existing health disparities and limits access to valuable resources to a thin and privileged slice of the population. Policy implications from the Office of Disease Prevention and Health Promotion report (2008) suggest that there is an urgent need to address the gap between publicly available health information and existing realities in health literacy levels across various socio-demographic spheres.

The importance of accessibility to comprehensible text becomes even more apparent considering that individuals with low health literacy are at higher risk for poorer access to care, experience poorer health outcomes (Berkman, Sheridan, Donahue, Halpern, & Crotty, 2011), and have higher hospitalization rates than individuals with high health literacy (McInnes & Haglund, 2011). According to a number of reports (Baker, Parker, Williams, & Clark, 1998; Baker et al., 2002; Gordon, Hampson, Capell, & Madhok, 2002; Scott, Gazmararian, Williams, & Baker,

2002), individuals with low health literacy make greater use of *treatment* services, as compared with services designed to *prevent* the onset of disease or lessen serious complications. This results in an estimated $50 - $73 billion dollars in additional health care costs annually in the United States. It is possible that one way to attenuate these costs might be to match the readability of written healthcare information to national reading grade level averages, or below. Although this is clearly not a catch-all strategy for reducing the financial burden associated with poor health outcomes, it is an important first step in addressing existing disparities in health literacy, and providing consumers with usable information from which they can make more informed decisions about their own, or loved ones', mental healthcare needs.

Readability

In accordance with this theme, several national organizations including the Centers for Disease Control and Prevention (CDC) and the American Medical Association (AMA) recommend that health information be written at a $6^{th} – 8^{th}$ grade reading level (Neuhauser & Paul, 2011; Weis, 2003). Grade level estimates of written text are synonymous with the concept of readability, which can be calculated in a number of ways. Typically, readability formulas give a general estimate of how difficult a text is to read based on the average number of syllables per word, and number of words per sentence. The readability score estimated from these formulas refers to the grade level people need to have completed to be able to read the text. It is important to remember that readability does not equate to comprehension, which can oftentimes be two or more grade levels below reading or education level, and drops when a person is under stress (McInnes & Haglund, 2011). Indeed, even individuals

with strong literacy skills and high educational attainment can face health literacy challenges, particularly when faced with: being diagnosed with a serious medical or mental illness that requires complicated self-care; unfamiliarity with opaque medical terminology and processes; and/or having to interpret numbers or risks in order to make challenging healthcare decisions (America's Health Literacy, 2008).

Common readability indices include the Automated Readability Index, Coleman Liau Index, Simple Measure of Gobbledygook (SMOG), Gunning Frequency of Gobbledygook (FOG) score, and Flesch-Kincaid Grade Level (Friedman & Hoffman-Goetz, 2006). These indices generate reading level scores based on unique formulas or algorithms, hence increasing the probability that scores obtained from each index will exhibit marked variability. Given that these five indices were used to assess the same sample of text for each disorder in this study, a modeling approach that takes into consideration clustering within the data was necessary in order to examine the relationship between website (source) and topic area (subject), when accounting for variability in reading grade level scores by index. This approach provides a robust method for assessing differences in readability scores between, and within, websites and content areas, respectively.

In summary, the purpose of this project was to systematically examine reading grade levels for 16 common mental health disorders from the top 6 websites common to all disorders. A significant source by content area interaction was hypothesized when accounting for the variability in reading level estimates generated by various indices, such that grade level estimates for various disorders were expected to vary based on the website text was derived from. It was also hypothesized that written text

for some of the more serious mental illnesses examined, such as schizophrenia, bipolar disorder, and borderline personality disorder, would have the highest reading level estimates, as compared with text from other disorders. Given the dearth of attention bestowed upon the readability of patient mental health materials in the past, it was expected that text from all websites would exceed the recommended 6th to 8th grade guidelines suggested by the CDC and the AMA.

Methods

Materials

According to the website ebizMBA.com, Google, Bing, Yahoo, Ask, and AOL.com, respectively, have been named the top five search engines of 2015. Because different Internet search engines may produce unique results for the same query based on numerous factors (including an individual's location and browsing history), top website hits for the sixteen disorders selected for analysis were explored using all five search engines. That is, each term of interest was entered using Google, Bing, Yahoo, Ask, and AOL, and the top fifteen website hits for each query were recorded and examined for consistency across search engines. This process of exploration and elimination resulted in the following list of six common websites that contain information for all disorders under investigation: Wikipedia.com, MayoClinic.org, PsychCentral.com, MedicineNet.com, HealthLine.com, and WebMd.com.

Information from the 'About Us' or 'About' tab on MayoClinic.org, PsychCentral.com, MedicineNet.com, HealthLine.com, and WebMd.com suggests that information on these sites is monitored and maintained by a team of editors, physicians, and other healthcare professionals. Indeed, as of July 14, 2016, information on the HealthLine.com site claims that "Health seekers have made us the fastest growing health information site. Over 40 million people turn to Healthline every month". Likewise, the MedicineNet.com site states that "MedicineNet is an online, healthcare media publishing company. We provide easy-to-read, in-depth, authoritative medical information for consumers via its robust, user-friendly,

interactive website. Founded in 1996, MedicineNet.com has had a highly accomplished, uniquely experienced team of qualified executives in the fields of medicine, healthcare, Internet technology, and business to bring you the most comprehensive, sought-after healthcare information anywhere. Nationally recognized, doctor-produced by a network of more than 70 U.S. board-certified physicians, MedicineNet.com and onhealth.com are trusted sources for online health and medical information". Despite being acquired by WebMd.com in 2004, MedicineNet.com maintains that it operates under this original vision. In light of this claim, and for the purposes of this study, information from MedicineNet.com was deemed independent from information obtained from WebMd.com.

Information obtained from the WebMd.com website indicates that "WebMD has created an organization that we believe fulfills the promise of health information on the Internet. We provide credible information, supportive communities, and in-depth reference material about health subjects that matter to you. We are a source for original and timely health information as well as material from well-known content providers". Information obtained from the MayoClinic.com website also indicates that "The product development team consists of experts in content development and production, product management, and user experience and design. Because physicians, scientists and other medical experts dedicate a portion of their clinical time to this site, we are in the unique position to give you access to the knowledge and experience of Mayo Clinic". Although the MayoClinic.com site provides detailed information about the mental health disorders and conditions explored in this study, as of July, 2016, the

site does not name any psychologists, psychiatrists, or licensed mental health workers under its list of specialty medical editors.

Information obtained from the PsychCentral.com website claims that its credo is to "Provide the best evidence-based mental health & psychology information, regardless of profession. All voices are important and should be elevated in the discourse about mental illness & mental health", and that "Psych Central is the Internet's largest and oldest independent mental health social network. Since 1995, our award-winning website has been run by mental health professionals offering reliable, trusted information and over 250 support groups to consumers. We are today's modern voice for mental health information, emotional support and advocacy. With the broadest online reach and recognition of any mental health network today, we touch the lives of over 7 million people around the world every month".

Unlike the other sites examined, Wikipedia.com is owned by the non-profit organization Wikimedia Foundation and is described on the site Wikipedia.com as "a free Internet encyclopedia that allows its users to edit almost any article accessible. Wikipedia is the largest and most popular general reference work on the Internet and is ranked among the ten most popular websites". Clearly, Wikipedia.com is not managed by a board of mental health professionals, and its users generate and edit most of the mental health content posted on the site. However, given its popularity, Internet users searching for medical and mental health conditions are often directed to this site for key information.

A selection of text from each website, for each disorder, was extracted and saved in a Word document as a simple text file during the last two weeks of October,

2015. All commas, quotation marks, apostrophes, hyperlinks, references, and headings were removed from the text, as specified by common guidelines for readability analysis (DuBay, 2004). All bulleted lists and sentence fragments followed by a colon or semicolon were also removed. The final word count for selected texts ranged from approximately 150 to 600 words; the average number of words per sentence ranged from approximately 10 to 30 words; and the average number of syllables per word ranged from 1.5 to 2.5.

Text was processed by pasting extractions into the appropriate field on the website read-able.com. This website is one of many free, online readability calculators, and generates five different grade-level readability estimates. Readability estimates derived from this site were cross-referenced with estimates from indices available on readability-score.com and readability-formulas.com. Specific indices examined included the Automatic Readability Index, Coleman Readability Index, SMOG, Gunning Fog Grade Level, and Flesch-Kincaid Grade Level Index.

The Flesch-Kincaid Grade Level and SMOG indices generate an approximate grade level score at which an average U.S. student in that grade can read the text. For example, a Flesch-Kincaid or SMOG score of 8.3 indicates that an average student in the eighth grade can read the text in question. Although the specific formulas for each index vary slightly, both scores are dependent on the number of syllables per word, and number of words per sentence in the text passage under investigation. Similarly, the Gunning Fog Grade level utilizes average sentence length and percentage of hard words (words that contain more than 3 syllables) to generate a grade level for written text, such that the ideal Fog score is a 7 or 8. Materials that receive a score of 10 are

considered hard, 15 are considered difficult, and 20 are considered very difficult (DuBay, 2004).

Unlike the aforementioned indices, the Coleman-Liau and Automated Readability Index generate a readability estimate that takes into consideration the number of characters per word, as well as the number of words per sentence. Hence, although each index employs a different mathematical formula to arrive at a grade level score, scores should largely be consistent across indices.

Statistical analyses

For the purposes of this analysis, each of the selected reading level indices served as a separate *rater* of the same excerpt of text. Hence, reading level scores were clustered by rater (index), with each rater examining a total of 96 excerpts of text, for sixteen disorders, from six different websites. Because we were not interested in exploring differences in reading level scores *between* raters, and the raters chosen were conceptualized as a representative selection of the entire body of available raters (reading level indices), a population-averaged or generalized estimating equations (GEE) approach was utilized to explore systematic differences between websites, content areas, and website by content area interactions on population averaged reading grade level scores.

GEE's are typically used to estimate population-average or marginal models that describe changes in the population mean of a given variable in relation to other important covariates, while also taking into account subject specific non-independence among observations (Hubbard et al., 2010). Although the authors considered using the mean score for all raters for each disorder to explore differences in reading level

scores across disorders and websites, this approach reduces the number of measurements in each subject cluster to one data point, which may reduce power. Hence, specific statistical methodologies, such as GEE, that accommodate correlations within clusters were considered more appropriate for the questions explored in this study.

Results

Preliminary Analyses. Data were analyzed using SAS Version 9.3 (Carey Institute, N.C.), and SPSS Version 21 (IBM, 2012). In order to determine the need for more complicated methodological techniques, the Intraclass Correlation Coefficient (ICC) was first calculated for rater (index). The ICC can be conceptualized as a general measurement of agreement or consistency between two or more raters or measuring methods, where a value of '1' represents perfect agreement, and a value of '0' represents no agreement at all. The purpose of this preliminary analysis was to determine the extent of variability in reading level scores attributable to differences in rating algorithms utilized by each index selected. Because we were primarily interested in exploring how reading level scores vary by website and content area, it was important to take this variability into careful consideration; evidence for variability by index would suggest a clustering effect in the data that would need to be accounted for in all subsequent analyses.

A two-way random effects model was specified for rater in order to assess variability in reading level scores between raters. A two-way random effects model was selected because the same indices were used to assess all selections of text, and the indices selected were chosen from a population of available indices used to calculate grade reading level scores. The ICC (2) assumes that the variance of the raters serves to add noise to any ratings obtained, and that the mean of rater error is zero. Results indicated that the estimated reliability between indices was 82.1%, with 95% CI [76.9, 86.6], using a consistency definition. As can be seen in Table 1 below, the mean for reading level scores generated by the Gunning Fog index was highest and

had the largest variability, whereas the mean for reading level scores generated by the SMOG index was lowest and had the smallest variability of the indices selected.

Overall, the indices selected were largely consistent in their ratings of readability across disorders and/or websites. Hence, it could be concluded that the indices chosen demonstrated sufficient consistency for further analysis. Given that the researchers were: 1) not interested in examining specific differences between raters (indices) across websites and disorders; and 2) wanted to increase power by retaining as much information as possible from the original dataset (collapsing the data by calculating a mean score for each disorder from each website would reduce the number of available data points from 480, with all raters considered separately, to 96 when scores are averaged), a GEE approach was utilized to account for any natural variation in outcomes attributable to rater specific effects.

Index	Mean	Std. Deviation
Auto Readability Index	11.766	2.5710
Coleman	14.523	1.6711
FK Grade Level	12.263	2.2784
Gunning Fog	15.625	2.5924
SMOG	11.377	1.8263

Table 1: Mean readability ratings by index

Main Analyses. Reading level values extracted from the websites sampled ranged from 1.3 to 21.5, with a mean of 13.07, and standard deviation of 2.85 (N = 480). Results from a one-sample t-test exploring differences between mean readability estimates obtained across all websites and disorders and the national 8th grade average suggest that the mean of the obtained sample is significantly higher than the national average, (mean difference = 5.04, $p < .001$, t =52.27, 95% CI of difference [4.85, 5.23]).

Results from the GEE suggest a significant website by content area interaction, $\chi^2(4, 480) = 192.57$, $p < .001$, when controlling for the presumed interdependencies between scores across indices. The main effects of subject, $\chi^2(4, 480) = 436.92$, $p < .001$, and website, $\chi^2(4, 480) = 1446.20$, $p < .001$, were also significant at the .05 level. Significance tests for all reported pairwise comparisons were adjusted using the Holm-Bonferroni method (Holms, 1979). See Figure 1 and Table 2 for specific details regarding pairwise comparisons.

Interaction Effects. *Specific phobia (8.7% prevalence).* Results for specific phobia suggest that text obtained from MedicineNet.com had the highest reading grade level estimate (grade level estimate = 16.10, 95% CI [14.59, 17.62]), as compared with estimates for text obtained from WebMd.com ($p < .001$, grade level estimate = 13.62, 95% CI [11.82, 15.42]), PsychCentral.com ($p < .001$, grade level estimate = 13.48, 95% CI [11.90, 15.07]), MayoClinic.com ($p < .001$, grade level estimate = 12.88, 95% CI [11.26, 14.51]), HealthLine.com ($p < .001$, grade level estimate = 11.16, 95% CI [9.44, 12.88]), and Wikipedia.com ($p < .001$, grade level estimate = 10.32, 95% CI [8.47, 12.17]), respectively. All comparisons with MedicineNet.com reached statistical significance at the .05 level. Reading level estimates were consistent with a mid-high school to college level reading level and well exceeded the recommended 6th to 8th grade reading level guidelines.

Substance abuse (8.2% prevalence). Results for substance abuse suggest that text obtained from Wikipedia.com had the highest reading grade level estimate (grade level estimate = 14.20, 95% CI [12.75, 15.66]), as compared with estimates for text obtained from PsychCentral.com ($p = 1.0$, grade level estimate = 12.96, 95% CI

[10.61, 15.31]), MedicineNet.com (p = 1.0, grade level estimate = 12.62, 95% CI [10.64, 14.60]), MayoClinic.com (p <.001, grade level estimate = 11.12, 95% CI [9.72, 12.52]), WebMd.com (p = .001, grade level estimate = 11.10, 95% CI [9.84, 12.36]), and HealthLine.com (p <.001, grade level estimate = 10.62, 95% CI [8.95, 12.29]), respectively. Only comparisons between Wikipedia.com and MayoClinic.com, WebMd.com, and HealthLine.com, respectively, reached a level of statistical significance at the .05 level. All grade level estimates exceeded the recommended 6th to 8th grade guidelines and were consistent with a high school to college age reading level.

Alcoholism (6.8% prevalence). Results suggest that content related to alcoholism derived from the MedicineNet.com site had the highest reading grade level estimate (grade level estimate = 16.46, 95% CI [14.86, 18.06]), as compared with WebMD.com (p = .011, grade level estimate = 13.42, 95% CI [11.91, 14.93]), Wikipedia.com (p <.001, grade level estimate = 11.2, 95% CI [9.83, 12.57]), HealthLine.com (p <.001, grade level estimate = 10.98, 95% CI [9.23, 12.73]), and MayoClinic.com (p <.001, grade level estimate = 10.18, 95% CI [9.11, 11.25]), in descending order. The difference in estimates between MedicineNet.com and PsychCentral.com (grade level estimate = 12.46, 95% CI [10.93, 13.99]) was not significant (p = .056, 95% CI [-.023, 8.02]); all other comparisons reached significance at the .05 level. However, no reading grade level estimate from the websites examined approached the suggested 6th -8th grade reading level.

Social phobia (6.8% prevalence). Results for social phobia suggest that text obtained from Wikipedia.com had the highest reading grade level estimate (grade

level estimate = 15.64, 95% CI [14.16, 17.12]), as compared with estimates for text obtained from MedicineNet.com (p = 1.0, grade level estimate = 15.10, 95% CI [13.04, 17.16]), MayoClinic.com (p = .315, grade level estimate = 14.78, 95% CI [13.24, 16.32]), WebMd.com (p <.001, grade level estimate = 14.44, 95% CI [12.87, 16.01]), PsychCentral.com (p <.001, grade level estimate = 12.6, 95% CI [11.39, 13.81]), and HealthLine.com (p <.001, grade level estimate = 11.8, 95% CI [9.86, 13.74]), respectively. Only comparisons between Wikipedia.com and WebMd.com, PsychCentral.com, and HealthLine.com, respectively, reached statistical significance at the .05 level. All estimates well exceeded the recommended reading level guidelines and were consistent with an advanced high school to college reading grade level.

Major depressive disorder (MDD) (6.7% prevalence). Results for MDD suggest that text obtained from Wikipedia.com had the highest reading grade level (grade level estimate = 15.04, 95% CI [13.61, 16.47], as compared with HealthLine.com (p = 1.0, grade level estimate = 14.8, 95% CI [13.06, 16.53]), WebMd.com (p = 1.0, grade level estimate = 14.6, 95% CI [12.88, 16.32]), MedicineNet.com (p <.001, grade level estimate = 14.18, 95% CI [12.60, 15.76]), MayoClinic.com (p <.001, grade level estimate = 11.76, 95% CI [10.22, 13.31]), and PsychCentral.com (p <.001, grade level estimate = 11.72, 95% CI [10.07, 13.37]), respectively. Scores obtained from Wikipedia.com were significantly different from scores obtained from MedicineNet.com, MayoClinic.com, and PsychCentral.com, respectively, but not for HealthLine.com or WebMd.com. All estimates for major depressive disorder exceeded the recommended guidelines and were consistent with a high school to college reading grade level.

Attention deficit and hyperactivity disorder (ADHD) (4.1% prevalence)

Examination of corrected pairwise comparisons suggests that for ADHD, the population averaged reading grade level estimate obtained from MedicineNet.com (grade level estimate = 15.56, 95% CI [13.61, 17.52]), was significantly higher (at the .05 level) than the estimate obtained from Wikipedia.com (p <.001, grade level estimate = 13.08, 95% CI [11.61, 14.55]), MayoClinic.com (p <.001, grade level estimate = 12.36, 95% CI [10.71, 14.01]), and WebMd.com (p <.001, grade level estimate = 8.90, 95% CI [7.21, 10.60]), but not for HealthLine.com (p = 1.00, grade level estimate = 14.36, 95% CI [12.96, 15.76]) and PyschCentral.com (p = .389, grade level estimate = 13.56, 95% CI [12.44, 14.68]). Overall, MedicineNet.com had the highest reading grade level estimates, followed by HealthLine.com, PsychCentral.com, Wikipedia.com, MayoClinic.com, and WebMD.com, respectively. All estimates obtained for ADHD exceeded the recommended 6th to 8th grade reading level guidelines set forth by the CDC and other similar public health organizations.

Post-traumatic stress disorder (PTSD) (3.5% prevalence). Results for PTSD suggest that text obtained from WebMD.com had the highest reading grade level estimate (grade level estimate = 15.44, 95% CI [13.99, 16.89]), as compared with estimates obtained from MedicineNet.com (p <.001, grade level estimate = 14.64, 95% CI [13.32, 15.96]), Wikipedia.com (p <.001, grade level estimate = 13.82, 95% CI [12.53, 15.11]), PsychCentral.com (p <.001, grade level estimate = 12.76, 95% CI [11.35, 14.18]), MayoClinic.com (p <.001, grade level estimate = 11.54, 95% CI [10.21, 12.87]), and HealthLine.com (p <.001, grade level estimate = 11.48, 95% CI [9.87, 13.09]), respectively. All comparisons with WebMd.com reached statistical

significance at the .05 level, exceeded the recommended guidelines, and were consistent with a high school and above reading grade level.

Generalized anxiety disorder (GAD) (3.1% prevalence). Examination of text obtained from MedicineNet.com related to generalized anxiety disorder revealed that MedicineNet.com had the highest reading grade level estimate (grade level estimate = 16.04, 95% CI [13.92, 18.16]), as compared with text obtained from Wikipedia.com (p = 1.0, grade level estimate = 15.3, 95% CI [13.76, 16.84]), MayoClinic.com (p <.001, grade level estimate = 12.86, 95% CI [10.61, 15.11]), HealthLine.com (p = .938, grade level estimate = 12.6, 95% CI [11.31, 13.89]), WebMd.com (p = .689, grade level estimate = 12.4, 95% CI [10.64, 14.16]), and PsychCentral.com (p <.001, grade level estimate = 11.78, 95% CI [10.21, 13.35]), respectively. Reading level estimates from MedicineNet.com were significantly higher than estimates obtained from MayoClinic.com and PsychCentral.com, but not from Wikipedia.com, HealthLine.com, or WebMd.com. All estimates obtained for generalized anxiety disorder were consistent with a high school or above reading grade level and exceeded the recommended 6^{th} to 8^{th} grade reading level guidelines.

Panic disorder (2.7% prevalence). Results for panic disorder suggest that text obtained from Wikipedia.com had the highest reading level estimate (grade level estimate = 14.74, 95% CI [13.33, 16.15]), as compared with estimates from MedicineNet.com (p <.001, grade level estimate = 12.52, 95% CI [11.30, 13.75]), WebMd.com (p <.001, grade level estimate = 12.3, 95% CI [11.07, 13.53]), HealthLine.com (p <.001, grade level estimate = 11.54, 95% CI [10.17, 12.91]), MayoClinic.com (p <.001, grade level estimate = 10.7, 95% CI [9.29, 12.11]), and

PsychCentral.com (p <.001, grade level estimate = 10.62, 95% CI [9.24, 12.00]), respectively. All comparisons reached statistical significance at the .05 level. Estimates from all websites for panic disorder exceeded the recommended 6th to 8th grade guidelines and were consistent with a high school to college age reading level.

Bipolar disorder (2.6% prevalence). Results suggest that reading grade level was highest for text derived from MedicineNet.com (grade level estimate = 16.56, 95% CI [14.77, 18.35]), as compared with text obtained from MayoClinic.com (p =.013, grade level estimate = 15.24, 95% CI [13.38, 17.1]), PsychCentral.com (p <.001, grade level estimate = 13.9, 95% CI [12.33, 15.47]), Wikipedia.com (p <.001, grade level estimate = 11.66, 95% CI [10.01, 13.31]), and HealthLine.com (p <.001, grade level estimate = 11.46, 95% CI [9.21, 13.71]). The difference between scores obtained from MedicineNet.com and WebMd.com (p = 1.00, grade level estimate = 15.68, 95% CI [13.66, 17.70]) was not significant. All estimates obtained for Bipolar disorder exceeded the recommended 6th to 8th grade reading level guidelines.

Borderline personality disorder (BPD) (1.6% prevalence). Results indicate that text related to borderline personality disorder extracted from the MedicineNet.com site had the highest reading grade level estimate (grade level estimate = 17.90, 95% CI [16.09, 19.71]), as compared with text obtained from HealthLine.com (p <.001, grade level estimate = 12.38, 95% CI [10.54, 14.22]), MayoClinic.com (p <.001, grade level estimate = 11.58, 95% CI [10.11, 13.05]), and WebMd.com (p <.001, grade level estimate = 9.32, 95% CI [7.58, 11.06]). The difference between scores obtained from MedicineNet.com and PsychCentral.com (p = 1.0, grade level estimate = 17.36, 95% CI [15.38, 19.34]), and MedicineNet.com and

Wikipedia.com (p = 1.0, grade level estimate = 16.4, 95% CI [14.67, 18.33]), was not significant. Overall, text obtained from MedicineNet.com had the highest reading grade level estimates, followed by text from PsychCentral.com, Wikipedia.com, HealthLine.com, MayoClinic.com, and WebMd.com, respectively. However, all estimates exceeded the recommended 6th to 8th grade reading level guidelines.

Schizophrenia (1.1% prevalence). Results for schizophrenia suggest that text obtained from MedicineNet.com had the highest reading grade level estimate (grade level estimate = 16.36, 95% CI [14.56, 18.16]), as compared with estimates for text obtained from PsychCentral.com (p <.001, grade level estimate = 14.66, 95% CI [13.12, 16.20]), WebMd.com (p <.001, grade level estimate = 13.54, 95% CI [11.83, 15.25]), Wikipedia.com (p <.001, grade level estimate = 13.48, 95% CI [11.74, 15.22]), HealthLine.com (p <.001, grade level estimate = 13.02, 95% CI [11.15, 14.89]), and MayoClinic.com (p <.001, grade level estimate = 13.02, 95% CI [11.55, 14.49]), respectively. All comparisons with MedicineNet.com reached statistical significance at the .05 level. Reading level estimates from all websites exceeded the recommended 6th to 8th grade guidelines and were consistent with a high school to college level and beyond reading level.

Obsessive compulsive disorder (OCD) (1.0% prevalence). Results for OCD suggest that text obtained from MedicineNet.com had that highest reading grade level (grade level estimate = 14.9, 95% CI [13.41, 16.39]), as compared with WebMd.com (p = 1.0, grade level estimate = 13.78, 95% CI [12.25, 15.31]), PsychCentral.com (p = 1.0, grade level estimate = 13.44, 95% CI [11.71, 15.18]), Wikipedia.com (p = .201, grade level estimate = 11.72, 95% CI [10.08, 13.36]), MayoClinic.com (p = .542,

grade level estimate = 11.66, 95% CI [9.74, 13.58]), and HealthLine.com (p = .005, grade level estimate = 11.0, 95% CI [9.16, 12.84]), respectively. Only the difference in reading level scores between MedicineNet.com and HealthLine.com was significant. All scores exceeded the recommended grade level guidelines and were consistent with an average high school and above reading level.

Agoraphobia (.8% prevalence). Examination of corrected pairwise comparisons suggests that for Agoraphobia, the population-averaged reading grade level estimate obtained from MedicineNet.com (grade level estimate = 16.54, 96% CI [14.93, 18.15]) was significantly higher (at the .05 level) than the estimate obtained for Wikipedia.com (p <.001, grade level estimate = 13.5, 95% CI [12.23, 14.77]), MayoClinic.com (p <.001, grade level estimate = 12.14, 95% CI [10.97, 13.31]), HealthLine.com (p <.001, grade level estimate = 11.16, 95% CI [8.85, 13.47]), and WebMd.com (p <.001, grade level estimate = 5.62, 95% CI [4.17, 7.07]), but not for PsychCentral.com (p = .251, grade level estimate = 12.42. 95% CI [10.81, 14.03]). Only information obtained from WebMd.com met the recommended reading level guidelines for printed health materials.

Bulimia nervosa (.3% prevalence). Exploration of results from Bonferroni-corrected pairwise comparisons for Bulimia nervosa indicate that text obtained from MedicineNet.com had the highest reading grade level estimate (grade level estimate = 15.02, 95% CI [13.55, 16.49]), followed by text obtained from WebMD.com (p = 1.0, grade level estimate = 14.98, 95% CI [13.60, 16.36]), PsychCentral.com (p = .048, grade level estimate = 13.96, 95% CI [12.50, 15.43]), MayoClinic.com (p <.001, grade level estimate = 12.04, 95% CI [10.59, 13.49]), HealthLine.com (p = .412, grade level

estimate = 11.82, 95% CI [9.90, 13.74]), and Wikipedia.com (p <.001, grade level estimate = 10.4, 95% CI [8.99, 11.81]), respectively. Only the difference in estimates between MedicineNet.com and MayoClinic.com, MedicineNet.com and Wikipedia.com, and MedicineNet.com and PsychCentral.com reached statistical significance at the .05 level. No estimates approached the recommended 6th – 8th grade reading level guidelines, and all estimates were consistent with a high school to college reading grade level.

Anorexia nervosa (.6% lifetime prevalence). For anorexia nervosa, results indicate that reading grade level was highest for text derived from MedicineNet.com (grade level estimate = 15.32, 95% CI [13.88, 16.76]), as compared with text derived from WebMD.com (p <.001, grade level estimate = 14.1, 95% CI [12.66, 15.54]), PsychCentral.com (p <.001, grade level estimate = 13.0, 95% CI [11.45, 14.55]), MayoClinic.com (p <.001, grade level estimate = 12.64, 95% CI [11.26, 14.02]), HealthLine.com (p <.001, grade level estimate = 11.58, 95% CI [9.67, 13.49]), and Wikipedia (p <.001, grade level estimate = 10.12, 95% CI [8.62, 11.62]), respectively. No reading grade level estimates from the websites examined approached the suggested 6th -8th grade reading level guidelines.

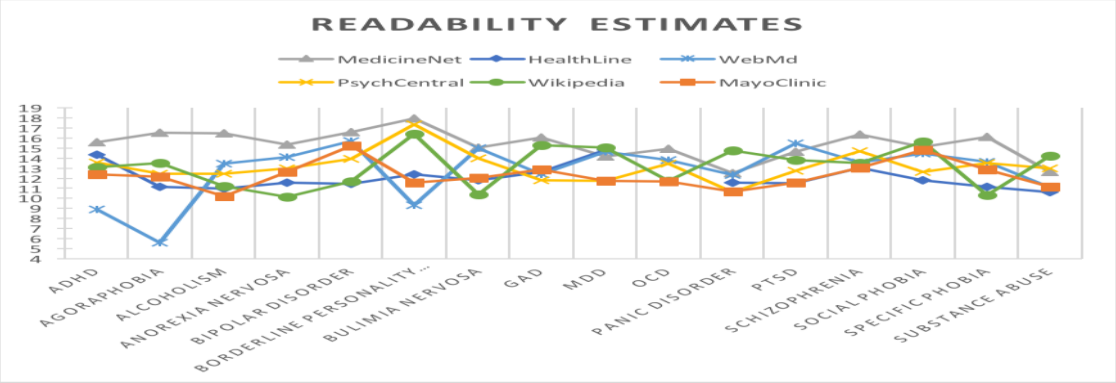

Figure 1: Readability estimates by website and disorder

Table 2: Reading Grade Level Estimates with 95% CI	*Estimate*	*Lower 95% CL*	*Upper 95% CL*
Agoraphobia	**5.62**	**4.17**	**7.07**
HealthLine	11.16	8.85	13.47
MayoClinic	12.14	10.97	13.31
MedicineNet	16.54	14.93	18.15
PsychCentral	12.42	10.81	14.03
WebMd	5.62	4.17	7.07
Wikipedia	13.5	12.23	14.77
ADHD	**8.9**	**7.21**	**10.6**
HealthLine	14.36	12.96	15.76
MayoClinic	12.36	10.71	14.01
MedicineNet	15.56	13.61	17.52
PsychCentral	13.56	12.44	14.68
WebMd	8.9	7.21	10.6
Wikipedia	13.08	11.61	14.55
Borderline Personality Disorder	**9.32**	**7.58**	**11.06**
HealthLine	12.38	10.54	14.22
MayoClinic	11.58	10.11	13.05
MedicineNet	17.9	16.09	19.71
PsychCentral	17.36	15.38	19.34
WebMd	9.32	7.58	11.06
Wikipedia	16.4	14.67	18.33
Anorexia Nervosa	**10.12**	**8.62**	**11.62**
HealthLine	11.58	9.67	13.49
MayoClinic	12.64	11.26	14.02
MedicineNet	15.32	13.88	16.76
PsychCentral	13	11.45	14.55
WebMd	14.1	12.66	15.54
Wikipedia	10.12	8.62	11.62
Alcoholism	**10.18**	**9.11**	**11.25**
HealthLine	10.98	9.23	12.73
MayoClinic	10.18	9.11	11.25
MedicineNet	16.46	14.86	18.06
PsychCentral	12.46	10.93	13.99
WebMd	13.42	11.91	14.93
Wikipedia	11.2	9.83	12.57
Specific Phobia	**10.32**	**8.47**	**12.17**
HealthLine	11.16	9.44	12.88

MayoClinic	12.88	11.26	14.51
MedicineNet	16.1	14.59	17.62
PsychCentral	13.48	11.9	15.07
WebMd	13.62	11.82	15.42
Wikipedia	10.32	8.47	12.17
Bulimia Nervosa	**10.4**	**8.99**	**11.81**
HealthLine	11.82	9.9	13.74
MayoClinic	12.04	10.59	13.49
MedicineNet	15.02	13.55	16.49
PsychCentral	13.96	12.5	15.43
WebMd	14.98	13.6	16.36
Wikipedia	10.4	8.99	11.81
Panic Disorder	**10.62**	**9.24**	**12**
HealthLine	11.54	10.17	12.91
MayoClinic	10.7	9.29	12.11
MedicineNet	12.52	11.3	13.75
PsychCentral	10.62	9.24	12
WebMd	12.3	11.07	13.53
Wikipedia	14.74	13.33	16.15
Substance abuse	**10.62**	**8.95**	**12.29**
HealthLine	10.62	8.95	12.29
MayoClinic	11.12	9.72	12.52
MedicineNet	12.62	10.64	14.6
PsychCentral	12.96	10.61	15.31
WebMd	11.1	9.84	12.36
Wikipedia	14.2	12.75	15.66
Bipolar Disorder	**11.46**	**9.21**	**13.31**
HealthLine	11.46	9.21	13.71
MayoClinic	15.24	13.38	17.1
MedicineNet	16.56	14.77	18.35
PsychCentral	13.9	12.33	15.47
WebMd	15.68	13.66	17.7
Wikipedia	11.66	10.01	13.31
PTSD	**11.48**	**9.87**	**12.87**
HealthLine	11.48	9.87	13.09
MayoClinic	11.54	10.21	12.87
MedicineNet	14.64	13.32	15.96
PsychCentral	12.76	11.35	14.18

WebMd	15.44	13.99	16.89
Wikipedia	13.82	12.53	15.11
OCD	**11.66**	**9.16**	**12.84**
HealthLine		9.16	12.84
MayoClinic	11.66	9.74	13.58
MedicineNet	14.9	13.4	16.39
PsychCentral	13.44	11.71	15.18
WebMd	13.78	12.25	15.31
Wikipedia	11.72	10.08	13.36
MDD	**11.72**	**10.07**	**13.31**
HealthLine	14.8	13.06	16.53
MayoClinic	11.76	10.22	13.31
MedicineNet	14.18	12.6	15.76
PsychCentral	11.72	10.07	13.37
WebMd	14.6	12.88	16.32
Wikipedia	15.04	13.61	16.47
GAD	**11.78**	**10.21**	**13.35**
HealthLine	12.6	11.31	13.89
MayoClinic	12.86	10.61	15.11
MedicineNet	16.04	13.92	18.16
PsychCentral	11.78	10.21	13.35
WebMd	12.4	10.64	14.16
Wikipedia	15.3	13.76	16.84
Social phobia	**11.8**	**9.86**	**13.74**
HealthLine	11.8	9.86	13.74
MayoClinic	14.78	13.24	16.32
MedicineNet	15.1	13.04	17.16
PsychCentral	12.6	11.39	13.81
WebMd	14.44	12.87	16.01
Wikipedia	15.64	14.16	17.12
Schizophrenia	**13.02**	**11.15**	**14.49**
HealthLine	13.02	11.15	14.89
MayoClinic	13.02	11.55	14.49
MedicineNet	16.36	14.56	18.16
PsychCentral	14.66	13.12	16.2
WebMd	13.54	11.83	15.25
Wikipedia	13.48	11.74	15.22

Main effects. *Subject.* Results suggest that the estimate for text related to borderline personality disorder had the highest reading grade level (grade level estimate = 14.157, 95% CI [12.75, 15.57]), as compared with all other disorders. The difference in estimates between borderline personality disorder and bipolar disorder (difference = .073, 95% CI [-1.01, 1.16]), social phobia (difference = .097, 95% CI [-.593, .787]), schizophrenia (difference = .143, 95% CI [-1.06, 1.35]), MDD (difference = .473, 95% CI [-.012, .959]), and GAD (difference = .660, 95% CI [-.138, 1.46]), respectively, was not significant at the .05 level. In summary, reading level estimates for borderline personality disorder, bipolar disorder, schizophrenia, MDD, and GAD ranged from 14.16 to 13.50; these estimates are well above the recommend guidelines and suggest an early college reading grade level.

Text related to borderline personality disorder was significantly higher in reading grade level as compared with text describing PTSD (difference = .877, 95% CI [.176, 1.58]), bulimia (difference = 1.12, 95% CI [.655, 1.59]), ADHD (difference = 1.19, 95% CI [.179, 2.20]), specific phobia (difference = 1.23, 95% CI [.402, 2.06]), anorexia (difference = 1.36, 95% CI [.784, 1.94]), OCD (difference = 1.41, 95% CI [.399, 2.41]), alcoholism (difference = 1.71, 95% CI [1.01, 2.40]), substance abuse (difference = 2.05, 95% CI [.964, 3.14]), panic disorder (difference = 2.09, 95% CI [1.26, 2.92]), and agoraphobia (difference = 2.26, 95% CI [-1.33, 3.19]). Reading level estimates for text related to these disorders ranged from 13.28 to 11.90; these estimates are well above the recommend guidelines and suggest an advanced high school to early college reading grade level.

Given the ranking of reading level estimates, and that there was a significant difference in estimates between borderline personality disorder and PTSD (the next highest-ranked disorder after GAD), pairwise comparisons were examined between PTSD and all remaining disorders in descending order of reading grade level. Results suggest that the difference in reading level estimates between PTSD (grade level estimate = 13.28, 95% CI [11.94, 14.64]) and bulimia (difference = .243, 95% CI [-.610, 1.10]), ADHD (difference = .310, 95% CI [-.447, 1.07]), specific phobia (difference = .353, 95% CI [-1.00, 1.71]), anorexia (difference = .487, 95% CI [-.501, 1.47]), and OCD (difference = .530, 95% CI [-.214, 1.27]), respectively, was not significant at the .05 level. Reading level estimates for these disorders ranged from 13.28 to 12.75 and are consistent with an early college reading grade level.

The reading level estimate for text describing PTSD was significantly higher than text describing alcoholism (difference = .830, 95% CI [.127, 1.53]), substance abuse (difference = 1.18, 95% CI [.140, 2.21]), panic disorder (difference = 1.21, 95% CI [.507, 1.91]), and agoraphobia (difference = 1.38, 95% CI [.084, 2.68]), respectively. Estimates for these disorders ranged from 12.45 to 11.90 and are consistent with an advanced high school/early college reading level.

Exploration of comparisons in reading level estimates by *type* of disorder revealed that social phobia had the highest reading grade level estimate (grade level estimate = 14.06, 95% CI [12.50, 15.62]) of all of the anxiety disorders examined, including GAD (difference = .563, 95% CI [.073, 1.05]), PTSD (difference = .780, 95% CI [.087, 1.47]), OCD (difference = 1.31, 95% CI [.759, 1.86]), panic disorder (difference = 1.99, 95% CI [1.44, 2.54]), and agoraphobia (difference = 2.16, 95% CI

[1.28, 3.04]). The difference between social phobia and specific phobia was not significant at the .05 level (difference = 1.13, 95% CI [-.058, 2.33]), nor was the difference between GAD and PTSD (difference = .217, 95% CI [-.595, 1.03]).

There was no difference in reading level estimates (difference = .347, 95% CI [-.264, .957]) between alcoholism (grade level estimate = 12.45, 95% CI [11.13, 13.77]), and substance abuse disorder (grade level estimate = 12.10, 95% CI [10.53, 13.68]). Likewise, there was no difference in reading level estimates (difference = .400, 95% CI [-.752, 1.55]) for the two mood disorders examined, bipolar disorder (grade level estimate = 14.08, 95% CI [12.32, 15.85]), and MDD (grade level estimate = 13.68, 95% CI [12.22, 15.15]). Estimates for these disorders were consistent with an advanced high school reading level and are well above the recommended 6th to 8th grade reading level guidelines for written patient health materials.

However, there was a significant difference (difference = .243, 95% CI [.022, .464]) in reading level estimates for text describing bulimia nervosa (grade level estimate = 13.04, 95% CI [11.62, 14.45]), and anorexia nervosa (grade level estimate = 12.79, 95% CI [11.31, 14.28]). Both estimates are consistent with an advanced high school to early college reading grade level.

Table 3: Mean readability estimates by subject

Subject	**Mean**	**Std. Error**	**95% Wald Confidence Interval**	
			Lower	*Upper*
ADHD	12.970	.7439	11.512	14.428
Agoraphobia	11.563	.9060	9.788	13.339
Alcoholism	12.450	.6727	11.131	13.769
Anorexia Nervosa	12.793	.7565	11.311	14.276
Bipolar	14.083	.9004	12.319	15.848
Borderline Personality Disorder	14.157	.7194	12.747	15.567

Bulimia Nervosa	13.037	.7221	11.621	14.452
GAD	13.497	.8001	11.929	15.065
MDD	13.683	.7482	12.217	15.150
OCD	12.417	.9456	10.563	14.270
Panic Disorder	12.070	.6590	10.778	13.362
PTSD	13.280	.6921	11.924	14.636
Schizophrenia	14.013	.8427	12.362	15.665
Social Phobia	14.060	.7963	12.499	15.621
Specific Phobia	12.927	.8243	11.311	14.542
Substance Abuse	12.103	.8020	10.531	13.675

Website. Overall, estimates for text obtained from MedicineNet.com had the highest reading level estimates of all websites examined (grade level estimate = 15.36, 95% CI [13.85, 16.88]), including PsychCentral.com (difference = 2.20, 95% CI [1.56, 3.53]), Wikipedia.com (difference = 2.20, 95% CI [1.40, 3.00]), WebMd.com (difference = 2.66, 95% CI [1.80, 3.53], MayoClinic.com (difference = 3.08, 95% CI [2.35, 3.82]), and HealthLine.com (difference = 3.38, 95% CI [2.07, 4.69]), respectively. The difference in estimates between PsychCentral.com (grade level estimate = 13.17, 95% CI [11.75, 14.59]), and Wikipedia.com (grade level estimate = 13.16, 95% CI [11.71, 14.62]), was not significantly different (difference = .004, 95% CI [-.419, .426]), nor was the difference in estimates between Wikipedia.com and WebMd.com (difference = .461, 95% CI [-.143, 1.07]). Overall, all grade level estimates well exceeded the recommended reading level guidelines suggested by the CDC and other similar organizations, and were consistent with an average advanced high school to advanced college reading grade level.

Table 4: Mean readability estimates by website

Website	Mean	Std. Error	95% Wald Confidence Interval	
			Lower	*Upper*
HealthLine	11.985	.8381	10.342	13.628
MayoClinic	12.281	.7384	10.834	13.728
MedicineNet	15.364	.7744	13.846	16.882
PsychCentral	13.043	.7910	11.492	14.593
WebMd	12.578	.7812	11.046	14.109
Wikipedia	13.164	.7403	11.713	14.615

Discussion

Overall, aside from a key few instances, the reading grade level for all disorders across the websites examined far exceeded the suggested 6th to 8th grade reading level guidelines established by the CDC and other similar organizations. In some cases, (i.e. text related to borderline personality disorder from MedicineNet.com), the estimated reading grade level reached as high as 17.9. This estimate suggests that, on average, only individuals with an advanced graduate degree (grade 17.9) would be able to read the selected text effectively. In other instances, (i.e. text related to ADHD and Agoraphobia from WebMd.com), reading grade level estimates were much lower, and consistent with a 6th to 8th grade reading level, respectively. These estimates suggest that an individual who completed the 6th to 8th grade could effectively read the selected text. However, all other estimates obtained were markedly higher, with a minimum average high school reading level required to effectively read the selected text.

Interestingly, text related to borderline personality disorder demonstrated the highest reading grade level estimate, followed by text related to bipolar disorder, social phobia, schizophrenia, MDD, and GAD, in descending order of grade level. Examination of estimates for these disorders generally suggests that an individual with an average post-high school reading level could effectively read the segments of text selected for analysis. Given the severity of impairment often associated with these disorders (particularly borderline personality disorder, bipolar disorder, and schizophrenia), it could be surmised that the information available online from the websites surveyed is not only relatively inaccessible to most healthy consumers, but

particularly to those struggling with serious mental illness. Indeed, as noted by Revheim et al., (2014), individuals with schizophrenia commonly display severe deficits in reading ability. Likewise, given impairments in reading ability among individuals with serious mental illness, Rotondi et al. (2007) suggest that most online sources of mental health information are not well-suited to the needs of this population.

Not surprisingly, little difference was noted in reading grade level estimates between MDD and bipolar disorder, as these disorders may share a common language regarding general symptoms of depression. Likewise, given similarities in language, symptom presentation, and etiology, there was no notable difference in reading level scores for alcoholism and substance abuse, as well as social phobia and specific phobia. However, this rationale could not be extended to text describing the two predominant eating disorders examined in this study: reading level estimates for bulimia nervosa were significantly higher than those for anorexia nervosa. It is possible that further exploration of text *content* may reveal emphasis on different features, symptoms, or etiology of each disorder, hence contributing to differences in reading level estimates.

Indeed, the reader is encouraged to recall that this study only examined the *readability* of online public mental health materials, and did not explore the content (or meaning) of text extracted from the sites selected. Readability is an important first component in understanding whether the structure and form of written material is largely digestible by the average reader. Based on national statistics that suggest the reading grade level of the average American citizen is between the 6^{th} to 8^{th} grade

(Kutner et al., 2006; Paasche-Orlow et al., 2005), materials describing mental health conditions, symptoms, and disorders that exceed this threshold may not be useful in helping the general population make important decisions about their own, or loved ones' healthcare needs.

It is also vital to remember that although readability is an important first element in broadly distinguishing the level of education required to read a passage of text, reading *comprehension* is oftentimes two to three grade levels *below* an individual's overall level of education or established reading grade level. This effect may further be exacerbated when an individual is under duress or struggling with a serious mental illness (McInnes & Hagland, 2011). As such, for the 77 million Americans with limited health literacy (America's Health Literacy, 2008; Kutner et al., 2006; Paasche-Orlow et al., 2005), much of the current mental health materials available online may be both unreadable and incomprehensible. This can have broad implications for perpetuating health disparities by limiting access to publicly available mental health information to a small segment of the population who already possess above average health literacy, have better access to resources, and consequently, may have better health outcomes than those with low health literacy.

Examination of reading level estimates by website suggests that on average, MedicineNet.com has the highest reading grade level, followed by PsychCentral.com, Wikipedia.com, WebMd.com, MayoClinic.com, and HealthLine.com, in descending order. There was no difference in reading level scores between PsychCentral.com and Wikipedia.com, and betweenWikipedia.com and WebMd.com. However, PsychCentral.com had higher overall reading grade level estimates than WebMd.com,

whereas estimates from WebMd.com were higher than those obtained from MayoClinic.com and HealthLine.com. No difference was noted between MayoClinic.com and HealthLine.com. Reading grade level estimates for all websites were consistent with a high school senior reading level or above. These results provide valuable evidence that online information, procured from the most popular health-related websites, for 16 of the most prevalent mental health disorders and/or conditions is written at a level that far exceeds the national reading grade level average. Writers of public mental health materials are well-advised to take great care in ensuring that the information provided to consumers is not only accurate, but also written in a manner that does not enhance existing health disparities by limiting access to knowledge to an already educated minority. Although it is most likely that this oversight is largely un- intentional (and can perhaps be tentatively attributed to a combination of factors including the global level of education of those writing public health materials, and/or a general lack of knowledge /awareness of statistics related to health literacy levels in the United States), failure to adhere to these guidelines can have broad public health implications (American's Health Literacy, 2008).

Lastly, it is important to consider the practical and methodological limitations of this study before making sweeping conclusions about the content of online public mental health materials. Clearly, individuals have a multitude of ways of arriving at the websites and disorders examined within the scope of this study. In many cases, searching for mental health information may begin by entering key words related to symptoms, rather than names of formal diagnoses. This study did not assess the mechanism by which people arrive at the websites selected, with the implied

understanding that based on common search terms, people will eventually be funneled to a web page describing a disorder whose symptoms are consistent with their initial search terms.

Furthermore, this study is in no way a comprehensive review of all mental health diagnoses, nor does it sample all websites with available online mental health materials. The websites selected for analysis were chosen, in part, because they contain information specific to each disorder under investigation. Some prominent mental health websites, such as the National Institute of Mental Health (NIMH.NIH.gov) were not selected because they did not promote information specific to substance abuse disorders or alcoholism. Likewise, given the speed at which technology changes, it is possible that the search engines selected in October, 2015 to conduct the initial investigation are no longer the most popular engines available.

From a methodological perspective, it may have been more robust to assess each block of text using additional readability indices, as well as to have multiple researchers select, clean, and process each block of text for enhanced inter-rater reliability. Although the researcher attempted to employ rigorous standards in selecting text for each disorder, it is possible that the selections may exhibit some bias. However, despite these limitations, this study provides some initial evidence that current readability estimates for 16 of the most prevalent mental health disorders common to all sites surveyed are well above the 6th to 8th grade reading level guidelines suggested by the CDC and AMA. This information is important for researchers interested in conducting more rigorous explorations of online mental health materials, policy makers interested in decreasing health disparities amongst

various socio-demographic groups, and editors of mental health websites dedicated to providing consumers with quality written health materials. Future directions for this work may include examination of online information for all existing mental health diagnoses, exploration of quality of content of written text, experimental manipulations of text with consumers in the laboratory, and/or evaluation of differences in comprehension for online information presented in written, versus auditory or interactive format.

References

Albright, J., de Guzman, C., Acebo, P., Paiva, D., Faulkner, M., & Swanson, J. (1996). Readability of patient education materials: implications for clinical practice. *Applied Nursing Research, 9*(3), 139-143.

America's Health Literacy: Why We Need Accessible Health Information. An Issue Brief from the U.S. Department of Health and Human Services (2008). Retrieved 7/15/2016 from: http://health.gov/communication/literacy/issuebrief/

Baker, D.W, Parker, R.M., Williams, M.V., & Clark, W.S. (1998). Health literacy and the risk of hospital admission. *Journal of General Internal Medicine, 13*(12), 791-798.

Baker, D.W., Gazmararian J.A., Williams, M.V., Scott, T., Parker, R.M., Green, D., Ren, J., & Peel, J. (2002). Functional health literacy and the risk of hospital admission among Medicare managed care enrollees. *American Journal of Public Health, 92*(8): 1278-1283.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low health literacy and health outcomes: an updated systematic review. *Annals of Internal Medicine, 155*(2), 97-107.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., Viera, A., Crotty, K., ... & Viswanathan, M. (2011). Health literacy interventions and outcomes: an updated systematic review.

Brigo, F., Otte, W.M., Igwe, S.C., Tezzon, F., & Nardone, R. (2015). Clearly written, easily comprehended? The readability of websites providing information on epilepsy. *Epilepsy and Behavior, 44*, 35 – 39.

Center for Behavioral Health Statistics and Quality (2015). *Behavioral health trends in the United States: Results from the 2014 National Survey on Drug Use and Health* (HHS Publication No. SMA 15-4927, NSDUH Series H-50). Retrieved from http://www.samhsa.gov/data/

Colaco, M., Svider, P. F., Agarwal, N., Eloy, J. A., & Jackson, I. M. (2013). Readability assessment of online urology patient education materials. *The Journal of Urology*, *189*(3), 1048-1052.

DuBay, W. H. (2004). The Principles of Readability. *Online Submission*.

Fox, S., & Duggan, M. (2013). Health Online 2013. Pew Research Center. Retrieved August, 2015 from http://www.pewinternet.org/2013/01/15/health-online-2013/

Friedman, D. B., & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, *33*(3), 352-373.

Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *The journal of Neuroscience*, *30*(32), 10601-10608.

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: what are the differences?. *Statistics in medicine*, *28*(2), 221-239.

Gordon, M.M., Hampson, R., Capell, H.A., & Madhok, R. (2002). Illiteracy in rheumatoid arthritis patients as determined by the Rapid Estimate of Adult Literacy (REALM) score. *Rheumatology, 41*(7): 750-754.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65-70.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*(4), 467-474.

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Kindig, D. A., Panzer, A. M., & Nielsen-Bohlman, L. (Eds.). (2004). *Health Literacy: A Prescription to End Confusion*. National Academies Press.

Kutner, M., Greenburg, E., Jin, Y., & Paulsen, C. (2006). The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. *National Center for Education Statistics*.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 13-22.

McEnteggart, G. E., Naeem, M., Skierkowski, D., Baird, G. L., Ahn, S. H., & Soares, G. (2015). Readability of Online Patient Education Materials Related to IR. *Journal of Vascular and Interventional Radiology*.

McInnes, N., & Haglund, B. J. (2011). Readability of online health information: implications for health literacy. *Informatics for health and social care*, *36*(4), 173-189.

Misra, P., Agarwal, N., Kasabwala, K., Hansberry, D. R., Setzen, M., & Eloy, J. A. (2013). Readability analysis of healthcare-oriented education resources from the American academy of facial plastic and reconstructive surgery. *The Laryngoscope*, *123*(1), 90-96.

National Institute of Mental Health, Statistics (2016). Retrieved 7/15/2016 from: http://www.nimh.nih.gov/health/statistics/index.shtml

National Institute on Alcohol Abuse and Alcoholism, Alcohol Facts and Statistics (2016). Retrieved 7/15/2016 from: http://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-facts-and-statistics

Neuhauser, L., & Paul, K. (2011). Readablity, comprehension, and usability. *Communicating risks and benefits: An evidence-based user's guide,* 129-148.

Paasche-Orlow, M. K., Parker, R. M., Gazmararian, J. A., Nielsen-Bohlman, L. T., & Rudd, R. R. (2005). The prevalence of limited health literacy. *Journal of General Internal Medicine, 20*(2), 175-184.

Perrin, A., & Duggan, M. (2015). Americans' Internet Access: 2000-2015. Pew Research Center. Retrieved August, 2015 from http://www.pewinternet.org/files/2015/06/2015-06-26_internet-usage-across-demographics-discover_FINAL.pdf

Revheim, N., Corcoran, C.M., Dias, E., Hellmann, E., Martinez, A., Butler, P.D., … & Javitt, D.C. (2014). Reading deficits in schizophrenia and individuals at high clinical risk: Relationship to sensory function, course of illness, and psychosocial outcome. *American Journal of Psychiatry, 171,* 949-959.

Rotondi, A.J., Sinkule, J., Haas, G.L., Spring, M.B., Litschge, C.M., Newhill, C.E., …& Anderson, C.M. (2007). Designing websites for persons with cognitive deficits: Design and usability of a psychoeducational intervention for persons with severe mental illness. *Psychological Services, 4*(3), 202-224.

Substance Abuse and Mental Health Services Administration (2014). *Results from the 2013 National Survey on Drug Use and Health: Summary of National Findings*, NSDUH Series H-48, HHS Publication No. (SMA) 14-4863. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Svider, P. F., Agarwal, N., Choudhry, O. J., Hajart, A. F., Baredes, S., Liu, J. K., & Eloy, J. A. (2013). Readability assessment of online patient education materials from academic otolaryngology–head and neck surgery departments. *American journal of otolaryngology*, *34*(1), 31-35.

Weiss, B. D. (2003). Health literacy. *A manual for clinicians. Chicago: American Medical Association Foundation and American Medical Association*.

WHO. 7th Global Conference on Health Promotion: track themes: Health literacy and health behavior. (2009). Retrieved August 2015 from: http://www.who.int/healthpromotion/conferences/7gchp/track2/en/

MANUSCRIPT- 2

To be submitted to the journal of Evaluation and the Health Professions

A methodological inquiry into strategies for evaluating written online mental health information: Suggestions for nested data

Dorothy D. Skierkowski, MA[1], Paul Florin, PhD[1], Lisa Harlow, PhD[1], Jason Machan, PhD[2], Minsuk Shim, PhD[3], Yinjiao Ye, PhD[4], Leslie Mahler, PhD[5]

[1]Department of Psychology, University of Rhode Island, Kingston, RI, USA

[2]Director of Biostatistics Core, Lifespan Corporation, Rhode Island Hospital, Providence, RI, USA

[3]Department of Education, University of Rhode Island, Kingston, RI, USA

[4]Department of Communication Studies, University of Rhode Island, Kingston, RI

[5]Department of Communication Disorders, University of Rhode Island, Kingston, RI

Corresponding Author: Dorothy D. Skierkowski, MA

Department of Psychology

Community Research and Services Team

University of Rhode Island

80 Washington Street

Providence, RI, 02903

Phone: +1-860-692-8209

Email: dskierkowski@uri.edu

Abstract

For many years, the Internet has provided people with a plethora of valuable health information. Likewise, the Internet has also served as a valuable resource for understanding how, why, when, and from what sources consumers most often seek out information about health-related topics. Much of these data are freely available online; however, little attention has been paid to how to best analyze these data. The purpose of this paper is to provide public health researchers with a concise, and easy to understand guide to the best possible methods for analyzing publicly available data sourced from the Internet. Specifically, different methods for analyzing readability estimates of text derived from sixteen different mental health disorders, extracted from six highly utilized mental health websites, will be discussed. This example will demonstrate the importance of considering how data are structured, particularly when there is evidence of, or a strong theoretical rationale for clustering within the data. Different interpretations of the Intraclass Correlation Coefficient (ICC) will be presented within this context, and modeling approaches that account for within cluster correlation (i.e., mixed modeling and generalized estimating equations) will be discussed in greater detail. Overall, the researchers hope to provide public health researchers with a valuable toolkit for better understanding how data sourced online can be analyzed effectively, without excessive technical jargon. Researchers interested in more technical explanations are referred to the reference section of this report.

Introduction

The readability of written patient health materials is a topic of great importance for public health researchers. Since the development and expansion of the Internet, much attention has been devoted to understanding how people utilize this resource to obtain health-related information (see for example, Baker, Wagner, Singer & Bundorf, 2003; Diaz, Griffith, Ng, Reinert, Friedmann, & Moulton, 2002; Eysenbach & Kohler (2002); and McMullan, 2006). Given that much of the information available online is in written text format, analysis of the readability of online materials is paramount in ensuring that information intended for consumers remains accessible to the average reader (see McInnes & Haglund (2011), Neuhauser & Pace (2011), and Weis (2003) for a full discussion of how the construct of readability is defined and related to reading comprehension). In the United States, recent estimates suggest that the average adult reads at the 6^{th} to 8^{th} grade reading level (Kutner, Greenburg, Jin, & Paulsen, 2006). This has profound implications for writers of public health information, as information presented at a reading level much higher than the national average has the potential to maintain or exacerbate existing health disparities by catering only to those consumers with a high reading and educational status. These effects may be particularly egregious for individuals experiencing stress or mental health concerns.

Luckily, much as the Internet is a valuable resource for consumers of health information, it is also a vast repository of publicly available data for researchers interested in evaluating the readability of online health and mental health information. This study demonstrates how these data can be analyzed using methods that account

for natural clustering by website, subject area, and/or readability index utilized to rate the text. In all cases, it is important to bear in mind how the research question of interest and interpretability of results may change in response to alternate conceptualizations for how data are structured within clusters. The purpose of this paper is thus two-fold: 1) to provide an overview of various methods for analyzing clustered data, including a discussion of the utility of the Intraclass Correlation Coefficient (ICC) and differences between fixed and random effects; and 2) to demonstrate how results may vary across two possible approaches to analyzing nested readability data from 6 different websites related to 16 different mental health disorders, using five separate readability rater indices.

Overview of methods for analyzing nested data. Numerous techniques for analyzing nested data are currently available using common computer programs such as SPSS (IBM Corp., Armonk NY) or SAS (SAS Institute, Cary, NC), among others. These methods range from more straightforward methods such as multilevel analysis (or hierarchical linear modeling) for cross-sectional data using single indicator and outcome variables, to multi-level mediation models involving multiple mediators and moderators. More complex analyses often include categorical or non-normal response data and modeling of longitudinal effects over multiple time-points.

Given this wide range in methods, choosing the appropriate analysis may seem like a daunting task. However, it is important to remember that study design and an emphasis on addressing key questions of interest are of primary concern in developing an appropriate data analytic plan. The methods described herein are presented as a sampling of the multitude of techniques available for the analysis of nested data, and

are discussed in order from the most ‘simple’ (in comparison with the other techniques discussed) to the most complex.

Fixed and random effects. Throughout this report, reference will be made to ‘fixed’ and ‘random’ effects in the context of multi-level modeling. Slight variations in the definition of fixed and random effects appear in the literature on mixed modeling depending on author orientation and intended message. According to Hamilton (2012), fixed effects typically refer to intercepts and slopes that are meant to describe the population as a whole, whereas random effects refer to intercepts and slopes that vary across subgroups within the sample. Within the Hierarchical Linear Modeling (HLM) framework, Warne et al. (2012) describe fixed effects as the average impact that an explanatory variable has on a dependent variable across all clusters or groups, and random effects as the degree of variation between clusters.

Likewise, Hayes (2006) describes random effects as effects that are allowed to vary between Level 2 (higher order) units, whereas fixed effects are those that have only a single value in the model for each Level 1 (lowest level) unit regardless of the Level 2 unit under which they are nested. Under the umbrella of ordinary regression analyses, the intercept and slope are both considered fixed effects, and the residual is considered random. In contrast, when accounting for nested data, it is possible to specify an intercept and slope for each Level 2 unit of the same predictor by setting some of the coefficients as random effects (Hayes, 2006). Overall, researchers have a high degree of flexibility in choosing how to specify fixed and random effects in the modeling process dependent upon their primary question of interest and research design.

In our readability example, multiple iterations are possible. The simplest approach might be to first calculate a readability score for each website and subject area combination by averaging the scores across all five raters. This results in one Level 2 predictor (website, i.e., Wikipedia, WebMd, etc.), and one Level 1 predictor (disorder, i.e., agoraphobia, anorexia nervosa, etc.). Disorders are conceptualized as being nested within websites, and the outcome variable of interest is average readability score (across all five raters). Conceptualizing disorders as individuals nested within different websites allows for exploration of: 1) the effect of the Level 2 predictor (website) on the outcome of interest (average readability score); 2) the effect of the Level 1 predictor (disorder, i.e. agoraphobia, anorexia nervosa, etc.) on the outcome of interest; and 3) any Level 1 by Level 2 interactions of interest (website by disorder interactions), all the while acknowledging the hierarchical nature of the data. However, given that this approach reduces the size of our sample from 480 units of analysis to 96, and that the data are limited in the number of Level 2 clusters (6 websites), it is likely that our power to detect an effect if one is present is limited. Furthermore, without significant theoretical rationale for conceptualizing how nesting occurs within the data, it is equally possible to conceptualize that websites are clustered within disorders.

Another possible conceptualization of the data posits that each readability score for each disorder, from each website, is independent of all other scores, regardless of the rater (index) it was derived from. Although this iteration retains all data, it assumes that scores generated using the same readability index are not related, and ignores possible correlations within the data attributable to rater-specific effects.

Rater contributions are ignored, and only the fixed effects of website and subject area are explored in the regression analysis.

This second conceptualization represents a naïve approach because it ignores the possibility that each index can be thought of as a unique rater of the same block of text that uses a distinct formula to generate readability scores. It could hence be argued that scores within raters are more similar to each other than scores between raters, and that there is thus a need to account for these dependencies in the modeling process. In the examples noted above, the similarities within raters are not accounted for either because an average readability score is calculated for each disorder by website combination (N decreases from 480 to 96), or because each rating is treated as independent of all others.

These conceptualizations are potentially problematic because either the total sample size is reduced, the number of Level 2 groups is small, and/or any interdependencies in the data are not explicitly accounted for. Alternatively, it is possible to retain all of the data by treating indices as 'individuals' who are making multiple ratings on various passages of text. Here, it is possible to explicitly account for similarities in rating strategies *within* individuals by conceptualizing that readability scores from distinct websites and content areas are nested within the five individual raters selected for this study. Within this framework, it is possible to not only retain all of the data, but also to account for interdependencies within the scores generated from the same raters. This can be accomplished in a number of ways.

First, using mixed modeling, an intercept-only random effects model can be specified with only a random intercept included for raters. This preliminary approach

allows researchers to calculate the ICC, or ratio of group-level variance over total variance, and determine the need for further nested modeling approaches. Here, the ICC represents the proportion of variance in the dependent variable that is explained by the grouping structure of the hierarchical model (Castro, 2002; Wears, R.L., 2008). Although some statistical references suggest that an ICC close to zero negates the need for multi-level or clustered data approaches, Hayes (2006) argues that values of the ICC as small as .05 can invalidate hypotheses tests and confidence intervals when clustering is not considered. In this context, a value of .05 would indicate that approximately 5% of the total variation in readability scores could be accounted for by which rater made the readability rating and thus the raters should be taken into account. More discussion on the ICC is given shortly to provide more input to researchers.

The researcher may then choose to add predictor variables to the model and explore how the ICC changes with each new addition. For instance, the researcher could include a random component for the rater variable, and specify the calculation of fixed effects for website and content area. This approach allows raters to vary on the mean of readability scores, but assumes that the degree of association between explanatory and outcome variables is the same for all raters. In this example, it is possible to determine the degree of variation in scores between raters, and account for this variation, if necessary. Likewise, the researcher is able to flexibly decide which coefficients are to be fixed, and which coefficients are allowed to vary based on theory and research design.

The mixed modeling approach described above might be particularly useful if we were interested in assessing differences between raters, had some theory or hypothesis concerning how scores between raters might vary, but assumed that the degree of association between the predictor variables and the dependent variable was the same for all raters included in the analysis. Likewise, if we hypothesized that the degree of association between our predictor(s) and readability scores varied between raters, we might specify a random component for the predictor(s) of interest. It is important to remember, that the more coefficients specified, the greater the cost in degrees of freedom. Because our sample size and the number of groups is relatively small, we may be limited to simpler methodological designs.

In contrast, if we were not interested in exploring differences *between* raters, but still wanted to account for the variability in readability scores due to rater effects, a population-averaged approach might be an appealing alternative. General(ized) Estimating Equations (GEE) provide one such flexible regression-based strategy. These models are appealing because: 1) they can handle a variety of correlated measure models, as well as a variety of outcome data (i.e., continuous, count, binary); and 2) are more flexible for missing data compared to other models (Zeger, Liang & Albert, 1988).

Although both approaches take variation in rater scores into consideration, there can be marked differences in how output from these analyses are interpreted, particularly when outcome data are binary or counts. For linear data, interpretation of estimates obtained using mixed modeling and GEE suggests that: coefficients derived from mixed modeling procedures represent the change in mean outcome for a unit

change in the associated grouping variable, keeping the random effect fixed; whereas coefficients derived from GEE represent the change in the mean outcome for a unit change in the associated grouping variable, across all levels of the grouping variable observed (Hubbard, et al., 2010).

Furthermore, whereas random-coefficient models typically explicitly address variation at both unit-specific and higher-order levels, GEE models assume simple random sampling of subjects representing a population, as opposed to a set of higher order groups. Hence, GEE models provide "population average" results and model the marginal expectation of the outcome variable as a function of the predictors specified. Interestingly, intercept-only random-coefficients linear mixed models generally produce the same estimates as those obtained from the exchangeable working correlation model in GEE, albeit with a difference in degrees of freedom. Here, equal variances for all observations and equal covariance of all possible paired observations within the statistical unit are assumed, as well as no correlation of observations made on different units (Hubbard et al., 2010).

Although there are numerous costs and benefits to each modeling strategy, fundamentally, the decision to employ GEE over mixed modeling (or vice versa) can be pared down to the researchers' primary question(s) of interest. If the objective was to make comparisons between the grouping variable and the outcome of interest, a mixed modeling approach might be better suited. However, if the goal was to account for variation in the outcome variable due to clustering within the data, but not to make direct comparisons between clusters, a GEE approach might be more applicable. In the latter instance, the researcher is modeling the marginal expectation of the outcome of

interest across all clusters, and assumes that subjects are drawn from a sample representing the population. For a more detailed technical explanation, including thorough discussion of assumptions relevant to both modeling strategies, the reader is referred to Hubbard et. al. (2010).

Intraclass Correlation. A discussion of clustered data analysis is not complete without detailed consideration of the ICC. One of the potential risks of using traditional statistical methods for analyzing clustered data is that estimated standard errors may be smaller than appropriate (Warne et al., 2012); this may result in increased probability for Type I error (Hox, 2010). The ICC is a quantitative measure of the degree of dependence in the data, such that it is possible to assess how similar subjects are to each other within clusters (Kenny, Kashy & Bolger, 1998; Peugh, 2010). The value of the ICC ranges from 0.0 (perfect independence) to 1.0 (all subjects are the same as others within the cluster) (Warne et al., 2012).

Traditionally, the ICC has been conceptualized as a measure of rater reliability, which is particularly relevant considering the conceptualization of the data used for the running example in this text (i.e., various readability indices as individual 'raters' of the same passage of text). In a seminal article on intraclass correlations, Shrout and Fleiss (1979) provide several examples of different uses for the ICC in the context of a reliability study of the ratings of several judges. The authors make the point that assessing whether judgments made by multiple observers are reliable is critical to knowing whether these measurements are meaningful. However, multiple forms of the ICC exist, and each is appropriate under a limited set of circumstances.

There are typically two ways of conceptualizing the ICC: the ICC (1) is a measure of the amount of variance in individual level responses attributable to group level properties, as described above; whereas the ICC also (2) is a measure of the reliability of group means (Castro, 2012). ICC (1) values are typically not affected by group size or the number of groups. However, because of slight variation in the formula used to calculate these coefficients, the ICC (2) *is* influenced by group size. Because ICC's are based on variance partitioning, they are subject to the same assumptions as analysis of variance (ANOVA), including homogeneity of variance, normality, and independence (Castro, 2002). In summary, the ICC provides an omnibus measure of dependency in the data, and can be used to determine the need for hierarchical or nested modeling procedures.

Hierarchical Linear Modeling (HLM). For multilevel analyses involving two levels (i.e., individuals nested within groups) HLM can generally be thought of as a two-step approach. The first step, or Level 1, typically involves estimating a separate regression for each group of interest with individual-level predictors and outcome. At Level 2, the variance in the Level 1 slopes and intercepts is modeled using the group-level variable. These equations are evaluated simultaneously (Castro, 2002; Diex-Roux, 2000; Luke, 2004). By treating clustered groups as their own level of data, as well as a combination of individual scores, it is possible to examine the cross-level influence of variables, thus developing a more nuanced and ecologically valid approach to examining real-world phenomenon, when theoretically applicable (Luke, 2004; Raudenbush & Bryk, 2002; Warne, et all, 2012).

HLM is a statistical procedure that uses maximum likelihood to estimate the variance components of Level 2 models. This technique assumes multivariate normality for variables. Other assumptions of HLM include that: Level 1 residuals are independent and normally distributed with a mean of zero and equal variances across groups; Level 1 predictors are independent of Level 1 residuals; random errors at Level 2 are multivariate normal and are independent among Level 2 units; the set of Level 2 predictors are independent of Level 2 residuals; and that Level 1 and Level 2 residuals are independent (Hofmann, 1997).

Model building in HLM is a multi-stage process, in which the researcher may consider three broad classes of models, starting with a null model with no Level 1 or Level 2 predictors (Luke, 2004). As noted above, this model may be useful for calculating the ICC and guiding further decision-making, and generally produces estimates equivalent to those obtained from the exchangeable working correlation model in GEE. Next, depending on the primary question of interest, the researcher might begin to add predictor variables into the model, allowing the intercept to vary for each identified cluster. The last class of models assumes variation in slopes *and/or* intercepts across Level 2 units, and can include interactions between individuals and group-level constructs.

As discussed, although the benefits of using HLM to model real-world phenomenon are plentiful, there are some important limitations of this approach that warrant further explication. Perhaps the most glaring of these limitations include: potential violations of the assumption of multivariate normality when considering cross-level interactions; restriction of the dependent variable to be operationalized at

the lowest level of analysis; and the need for fairly large sample sizes to obtain a sufficient level of power (Castro, 2002; Hofmann, 1997). In our example using readability data derived online, a multi-level or HLM approach using all of the data may not be the best approach given our conceptualization of the data as readability scores related to different disorders from different websites, nested within different raters selected from a population of all possible raters.

General(ized) Linear Mixed Modeling (GLMM). HLM is a powerful technique for analyzing continuous outcome data. However, the assumptions of HLM do not hold when the response format is binary, multinomial, a proportion, or a count. For instance, if we were interested in whether websites passed or failed a reading grade level standard, or the influence of various factors on the number of websites that scored at the average reading level (rather than a continuous readability outcome measure), other statistical methods that take into consideration non-normal response formats would be necessary. GLMM is an extension of linear mixed modeling procedures that can readily handle non-normal data. This is particularly important when considering that much of the data collected online, in hospitals, schools, or other naturalistic community settings may follow a variety of alternative distributions (i.e., Poisson, binomial, negative binomial, etc.), and that the assumptions of linearity, normality, and constant variance may thus not be applicable. As such, GLMM is acceptable for determining Level 1 and Level 2 effects for non-normal or non-linear data, hence allowing for multi-level analysis of binary, count, ordinal, and multinomial data.

Kaplan (2004) suggests that some additional steps that must be taken when estimating generalized linear mixed models. First, a sampling model and link function must be specified. The link function transforms the expected value into a predicted value that can be estimated with a linear equation. In the case of linear mixed modeling, this is a normal distribution with a mean and variance, and a link function with the value of 1 (because no transformation is required). For binary outcomes (Y = 1, N = 0), this would mean a Bernoulli distribution and a log odds ratio link function. Next, the researcher must specify a linear structural model to estimate the transformed expected value. Conditional models may be specified, such that the researcher has the option of including relevant Level 1 or Level 2 predictors, and including fixed and random effects, as needed (Kaplan, 2004).

Furthermore, when considering generalized linear mixed models, a distinction should be made between unit-specific and population average models (Raudenbush & Bryk, 2002). For instance, the unit-specific model (hierarchically structured model) describes processes that are occurring in each Level 2 cluster, where processes are captured by the beta-coefficients of the Level 1 model. Here, the primary question of interest may be how the processes differ over a population of Level 2 units. It may be possible that these processes differ in their intercept alone, slope, or both. Furthermore, the Level 2 model may also assess how differences in the Level 2 explanatory variables influence Level 1 processes in each Level 1 unit. Hence, unit-specific models provide information about how effects of predictors vary across groups (Kaplan, 2004; Raudenbush & Bryk, 2002).

Raudenbush (2000; 2004) describes these questions as 'unit-specific', and contrasts them to a population-average (or Generalized Estimating Equations) approach (Zeger, Liang & Albert, 1988), in which the primary question of interest is in estimating average probabilities for population-level effects. Given the complexity and flexibility of this approach, one limitation may be that GLMM requires that researchers be explicit about their research questions and the type of data available for analysis, a priori. Interestingly, in some ways this could be conceptualized as both a weakness *and* strength of this approach, largely because it forces the researcher to exert much time and effort into clearly delineating their specific research hypotheses or intended intervention effects.

Structural equation modeling. Over the past number of years, structural equation modeling (SEM) has been studied and applied as a valid methodology for the analysis of multilevel or clustered data (Tomarken & Waller, 2005). Indeed, one of the primary strengths of SEM is the ability to specify latent variable models that provide estimates of the associations between latent constructs and their indicators (otherwise known as the measurement model), as well as between important constructs themselves (the structural model).

Using this framework, it is possible to account for biases that are attributable to random error and variation that is not better explained by the constructs of interest. Other general strengths of SEM include the ability to evaluate complex models with a large number of linear equations against less complex models, as well as the ability to specify recursive relationships between constructs (and error terms), hence accounting

for dependencies in data that are nested or collected repeatedly on the same individuals over time.

In a comparison of HLM with SEM, Raudenbush & Bryk (2002) suggest a number of striking similarities between these modeling approaches for two-level 'growth models', in which repeated measurements are taken on the same individuals over time. Here, the authors indicate that the Level 1 model of HLM corresponds with the measurement model of SEM, and that the latent variables of SEM are the individual growth parameters of HLM. The Level 2 model thus corresponds with the structural model specified in SEM. Using an SEM approach, it is hence possible to include Level 1 autocorrelated or heterogeneous random effects, and test a wide range of covariance structures (Duncan & Duncan, 2009). Curran (2003) further supports this claim and indicates that there is a large body of literature demonstrating that SEM and multilevel modeling are essentially analytically and empirically equivalent methods for evaluating clustering due to measurement of repeated observations over time.

However, one downfall is that SEM typically requires balanced data within groups, such that each individual is required to have the same number and distance between time points. Furthermore, Level 1 predictors with random effects are required to have the same distribution across all cases within each group. Unlike SEM, the HLM framework allows for unequal group sizes and spacing of time points, and does not require the distributions of Level 1 random effects to be identical (Raudenbush & Bryk, 2002). In recent years, the SEM framework has been extended to analyze data beyond a latent growth curve format, such that it is now possible to use SEM to

examine clustered data in situations that do not involve repeated measurements (Heck & Thomas, 2015; Hox & Maas, 2001; Tomarken & Waller, 2005).

Some attention has also been focused on extending the assumptions of multi-level SEM to include non-linear response formats, such that it is now possible to model categorical or count data within the multi-level SEM framework (Rabe-Hesketh, Skronda & Pickles, 2004). Generalized linear latent and mixed modeling (GLLAMM) combines features of generalized linear mixed modeling with structural equation modeling to produce a flexible and unified modeling framework that is capable of: handling data missing at random and has the scope for handling data not missing at random; dealing with unbalanced multilevel designs; allowing random coefficients of unbalanced covariates; including regressions among factors and/or random coefficients (latent variables) that vary at different levels; and modeling of ordered and unordered categorical responses, counts, and a wide range of alternative responses processes (Rabe-Hesketh, Skrondal & Pickles, 2004).

Given these capabilities, it has become increasingly apparent that boundaries between HLM, GLMM, and SEM have become somewhat blurry, and that researchers are now faced with the important task of deciding which framework is best suited for their data and their most relevant research hypotheses (Tomarken & Waller, 2005). Indeed, a return to fundamental questions of interest in any research design can be a guiding beacon of light for those who find themselves bogged down in the murky waters of 'analysis paralysis' in search of the 'best' analytic method. It is important for researchers to remember that the 'best' modeling strategy is that which is most suited

to their research design, and that no strategy can ultimately save those who fail to thoroughly plan for their journey into unexplored research lands.

Analysis of clustered data and issues related to sample size. A discussion of the analysis of clustered data using the techniques described above also warrants some mention of concerns related to sample size. There is some consensus that group-level sample size is more important than total sample size, with some compensation for a small number of groups in large individual-level samples (Maas & Hox, 2005). In a simulation study of sufficient sample sizes for multi-level modeling, Maas & Hox (2005) indicate that a small sample size at Level 2 (less than 50 groups) can lead to biased estimates of the Level 2 standard errors. Hence, the researchers strongly suggest using caution when applying multi-level methods with a limited number of groups, and call for bootstrapping or other simulation methods to account for these concerns when analyzing small-sample data.

In light of these concerns, and the high probability that modeling real-world phenomenon often involves a small or limited number of Level 2 groups, Hoyle and Gottfredson (2015) make several recommendations for maximizing the yield of multi-level modeling or SEM efforts when N's are small. These recommendations include: retaining all cases where possible in the analysis sample, such that no data are left unmodeled; optimizing the observed data to achieve normality and using reliable measures; and fixing or constraining variables where possible using knowledge from previous research to decrease the number of parameters that need to be estimated.

Summary. After careful consideration of the key points discussed above, two modeling strategies for assessing the readability of online mental health materials

using the full dataset described herein stand out as distinct possibilities. First, the data could be conceptualized as following a 2-level hierarchy, with scores from various disorders and websites nested within the five raters selected for this analysis. However, because of the small number of higher-order groups, as well as the relatively small size of our sample, it is hypothesized that utilizing a 2-level multilevel modeling approach may not be advisable.

Second, we could conceptualize that the raters selected are a random sample of all possible raters of online material, and although we are not interested in addressing differences *between* raters, we *are* interested in accounting for clustering within the data. Given this important design consideration, a general estimating approach could likewise be considered because it is better suited to our primary question of interest (i.e. assessing differences in reading level scores between websites and disorders across the population of possible raters). Results from these approaches will be discussed herein, with an emphasis on demonstrating that GEE may be better suited to the structure of these data, as well as the underlying research question of interest. However, because the response format is linear, it is likewise expected that results will not vary widely between approaches, and that the fundamental consideration for researchers selecting an appropriate methodology for analyzing these types of data will be conceptual in nature.

Methods

Materials. According to the website ebizMBA.com, Google, Bing, Yahoo, Ask, and AOL.com, respectively, have been named the top five search engines of 2015. Because different Internet search engines may produce unique results for the same query based on numerous factors (including an individual's location and browsing history), top website hits for the sixteen disorders selected for analysis were explored using all five search engines. That is, each term of interest was entered using Google, Bing, Yahoo, Ask, and AOL, and the top fifteen website hits for each query were recorded and examined for consistency across search engines. This process of exploration and elimination resulted in the following list of six common websites that contain information for all disorders under investigation: Wikipedia.com, MayoClinic.org, PsychCentral.com, MedicineNet.com, HealthLine.com, and WebMd.com.

A selection of text from each website, for each disorder, was extracted and saved in a Word document as a simple text file during the last two weeks of October, 2015. All commas, quotation marks, apostrophes, hyperlinks, references, bulleted lists, sentence fragments followed by a colon or semicolon, and headings were removed from the text, as specified by common guidelines for readability analysis (DuBay, 2004). The final word count for selected texts ranged from approximately 150 to 600 words; the average number of words per sentence ranged from approximately 10 to 30 words; and the average number of syllables per word ranged from 1.5 to 2.5.

Text was processed by pasting extractions into the appropriate field on the website read-able.com and was cross-referenced with estimates from indices available

on readability-score.com and readability-formulas.com. Specific indices examined included the Automatic Readability Index, Coleman Readability Index, SMOG, Gunning Fog Grade Level, and Flesch-Kincaid Grade Level Index. Although each index employs a different mathematical formula to arrive at a grade level score, scores were expected to largely be consistent across indices because the selected indices all measure the same construct.

Results

ICC. A two-way random effects model was specified for rater in order to assess variability in reading level scores between raters. A two-way random effects model was selected because the same indices were used to assess all selections of text, and the indices selected were chosen from a population of available indices used to calculate grade reading level scores. The ICC (2) assumes that the variance of the raters serves to add noise to any ratings obtained, and that the mean of rater error is zero. Results indicated that the estimated reliability between indices was 82.1%, with 95% CI [76.9, 86.6], using a consistency definition. The mean for reading level scores generated by the Gunning Fog index was highest and had the largest variability, whereas the mean for reading level scores generated by the SMOG index was lowest and had the smallest variability of the indices selected.

Overall, the indices selected were largely consistent in their ratings of readability scores across disorders and/or websites. Calculation of the ICC using a definition of absolute agreement revealed that although the various raters selected were consistent in their scoring, and could be thought of as reliable raters of reading grade level, they were not in absolute agreement on ratings of readability scores, ICC (2) = .483, 95% CI [.156, .700]. This distinction between consistency and absolute agreement can be best explained using the following example: score sets of (2,4), (4,6), and (6,8) can be thought of as perfectly consistent (ICC = 1.0), however, are not in perfect absolute agreement. For our purposes, measuring the consistency of reading level scores across raters is important because it tells us that raters are largely in agreement over how scores are assessed. Here, we can be confident that although there

are differences in the scores generated by the raters selected, as a whole, they are largely consistent in how they measure grade level readability for the disorders and websites selected.

Alternatively, we could also use the ICC to determine the percentage of total variance in the outcome (readability score) that can be explained by the grouping variable (rater/index). Results from the unconditional intercept-only model (ICC = .409) suggest that approximately 41% of the total variation in reading level scores can be attributed to rater effects (i.e., which rater or index makes the rating). These results suggest that overall, consideration of rater effects, is important in the modeling process.

2-Level Multilevel Model. A two-level multilevel modeling approach was applied in order to assess the effects of website and content area on readability scores across the indices selected for this study. Results from the unconditional intercept-only model suggest that approximately 41% of the variance in reading level scores was attributable to differences between raters, as noted above. Main and interaction effects for website and content area were then added to the model, while accounting for variation within the grouping variable. This was accomplished by adding 'rater' as the subject variable, and specifying fixed effects for the website and content area variables. In other words, it was presumed that although there would be some variability in average readability scores across raters, the direction of the association between the explanatory variables and the outcome would largely be consistent. Results from this model suggest a significant website by content area interaction, $F(75, 380) = 12.76$, $p < .001$, with content area and website estimates and their 95%

confidence intervals presented in Table 1 below. Comparing with the results of the null model, adding the website and content area variables had little effect on the variance between raters, and reduced the amount of variance at the within-group level from 5.28 to 1.22. After accounting for the effects of website, content area, and their interaction, the total amount of unexplained variance due to a difference between raters decreased to 25.24%.

Table 1: Subject and content area estimated means, with 95% CI, from 2-level multi-level modeling analysis

Subject	**Website**	**Mean**	**95% Confidence Interval**	
			Lower Bound	Upper Bound
ADHD	*HealthLine*	14.360	12.427	16.293
	MayoClinic	12.360	10.427	14.293
	MedicineNet	15.560	13.627	17.493
	PsychCentral	13.560	11.627	15.493
	WebMd	8.900	6.967	10.833
	Wikipedia	13.080	11.147	15.013
Agoraphobia	*HealthLine*	11.160	9.227	13.093
	MayoClinic	12.140	10.207	14.073
	MedicineNet	16.540	14.607	18.473
	PsychCentral	10.420	8.487	12.353
	WebMd	5.620	3.687	7.553
	Wikipedia	13.500	11.567	15.433
Alcoholism	*HealthLine*	10.980	9.047	12.913
	MayoClinic	10.180	8.247	12.113
	MedicineNet	16.460	14.527	18.393
	PsychCentral	12.460	10.527	14.393
	WebMd	13.420	11.487	15.353
	Wikipedia	11.200	9.267	13.133
Anorexia Nervosa	*HealthLine*	11.580	9.647	13.513
	MayoClinic	12.640	10.707	14.573
	MedicineNet	15.320	13.387	17.253
	PsychCentral	13.000	11.067	14.933
	WebMd	14.100	12.167	16.033
	Wikipedia	10.120	8.187	12.053
Bipolar	*HealthLine*	11.460	9.527	13.393
	MayoClinic	15.240	13.307	17.173
	MedicineNet	16.560	14.627	18.493
	PsychCentral	13.900	11.967	15.833
	WebMd	15.680	13.747	17.613
	Wikipedia	11.660	9.727	13.593

Borderline personality disorder	*HealthLine*	12.380	10.447	14.313
	MayoClinic	11.580	9.647	13.513
	MedicineNet	17.900	15.967	19.833
	PsychCentral	17.360	15.427	19.293
	WebMd	9.320	7.387	11.253
	Wikipedia	16.400	14.467	18.333
Bulimia Nervosa	*HealthLine*	11.820	9.887	13.753
	MayoClinic	12.040	10.107	13.973
	MedicineNet	15.020	13.087	16.953
	PsychCentral	13.960	12.027	15.893
	WebMd	14.980	13.047	16.913
	Wikipedia	10.400	8.467	12.333
GAD	*HealthLine*	12.600	10.667	14.533
	MayoClinic	12.860	10.927	14.793
	MedicineNet	16.040	14.107	17.973
	PsychCentral	11.780	9.847	13.713
	WebMd	12.400	10.467	14.333
	Wikipedia	15.300	13.367	17.233
MDD	*HealthLine*	14.800	12.867	16.733
	MayoClinic	11.760	9.827	13.693
	MedicineNet	14.180	12.247	16.113
	PsychCentral	11.720	9.787	13.653
	WebMd	14.600	12.667	16.533
	Wikipedia	15.040	13.107	16.973
OCD	*HealthLine*	11.000	9.067	12.933
	MayoClinic	11.660	9.727	13.593
	MedicineNet	14.900	12.967	16.833
	PsychCentral	13.440	11.507	15.373
	WebMd	11.780	9.847	13.713
	Wikipedia	11.720	9.787	13.653
Panic Disorder	*HealthLine*	11.540	9.607	13.473
	MayoClinic	10.700	8.767	12.633
	MedicineNet	12.520	10.587	14.453
	PsychCentral	10.620	8.687	12.553
	WebMd	12.300	10.367	14.233
	Wikipedia	14.740	12.807	16.673
PTSD	*HealthLine*	11.480	9.547	13.413
	MayoClinic	11.540	9.607	13.473
	MedicineNet	14.640	12.707	16.573
	PsychCentral	12.760	10.827	14.693
	WebMd	15.440	13.507	17.373
	Wikipedia	13.820	11.887	15.753
Schizophrenia	*HealthLine*	13.020	11.087	14.953
	MayoClinic	13.020	11.087	14.953
	MedicineNet	16.360	14.427	18.293
	PsychCentral	14.660	12.727	16.593
	WebMd	13.540	11.607	15.473
	Wikipedia	13.480	11.547	15.413
Social Phobia	*HealthLine*	11.800	9.867	13.733
	MayoClinic	14.780	12.847	16.713
	MedicineNet	15.100	13.167	17.033

	PsychCentral	12.600	10.667	14.533
	WebMd	14.440	12.507	16.373
	Wikipedia	15.640	13.707	17.573
Specific Phobia	*HealthLine*	11.160	9.227	13.093
	MayoClinic	12.880	10.947	14.813
	MedicineNet	16.100	14.167	18.033
	PsychCentral	13.480	11.547	15.413
	WebMd	13.620	11.687	15.553
	Wikipedia	10.320	8.387	12.253
Substance Abuse	*HealthLine*	10.620	8.687	12.553
	MayoClinic	11.120	9.187	13.053
	MedicineNet	12.620	10.687	14.553
	PsychCentral	12.960	11.027	14.893
	WebMd	11.100	9.167	13.033
	Wikipedia	14.200	12.267	16.133

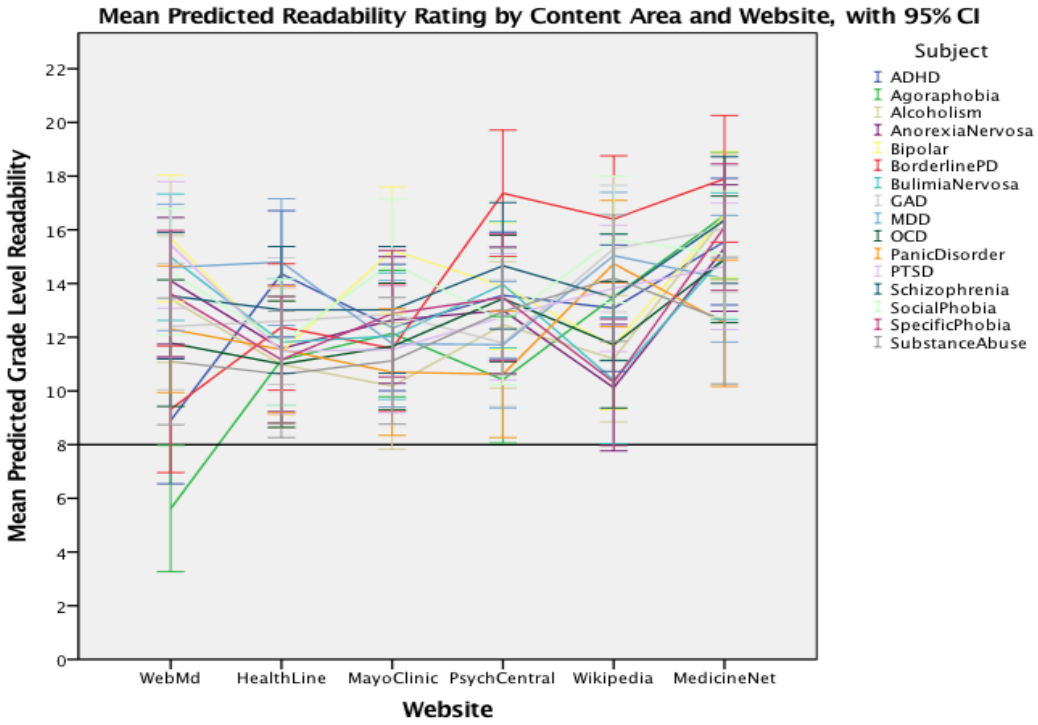

Figure 1: Mean predicted readability rating by content area and website, with 95% CI

GEE. Reading level values ranged from 1.5 to 21.5, with a mean of 13.11, and standard deviation of 2.77 (N = 480). Modeling these data using a GEE approach allowed the researchers to retain data from all of the indices examined, as well as to

account for any clustering in the data due to rater or index effects. Results from the GEE suggest a significant website by content area interaction, $\chi^2(4, 480) = 192.57$, $p < .001$, when controlling for the presumed interdependencies between scores across indices. The main effects of subject, $\chi^2(4, 480) = 572.80$, $p < .001$), and website, $\chi^2(4, 480) = 8376.32$, $p < .001$, were also significant at the .05 level. Holm-Bonferroni's correction was applied to adjust for multiple comparisons. Results from this analysis are presented in Table 2 below.

Table 2: Subject and content area estimated means, with standard error and 95% CI, from 2-level multi-level modeling analysis

Subject	**Estimate**	**Lower 95% *CL***	**Upper 95% *CL***
Agoraphobia			
HealthLine	11.16	8.85	13.47
MayoClinic	12.14	10.97	13.31
MedicineNet	16.54	14.93	18.15
PsychCentral	12.42	10.81	14.03
WebMd	5.62	4.17	7.07
Wikipedia	13.5	12.23	14.77
ADHD			
HealthLine	14.36	12.96	15.76
MayoClinic	12.36	10.71	14.01
MedicineNet	15.56	13.61	17.52
PsychCentral	13.56	12.44	14.68
WebMd	8.9	7.21	10.6
Wikipedia	13.08	11.61	14.55
Borderline Personality Disorder			
HealthLine	12.38	10.54	14.22
MayoClinic	11.58	10.11	13.05
MedicineNet	17.9	16.09	19.71
PsychCentral	17.36	15.38	19.34
WebMd	9.32	7.58	11.06
Wikipedia	16.4	14.67	18.33
Anorexia Nervosa			
HealthLine	11.58	9.67	13.49
MayoClinic	12.64	11.26	14.02

MedicineNet	15.32	13.88	16.76
PsychCentral	13	11.45	14.55
WebMd	14.1	12.66	15.54
Wikipedia	10.12	8.62	11.62
Alcoholism			
HealthLine	10.98	9.23	12.73
MayoClinic	10.18	9.11	11.25
MedicineNet	16.46	14.86	18.06
PsychCentral	12.46	10.93	13.99
WebMd	13.42	11.91	14.93
Wikipedia	11.2	9.83	12.57
Specific Phobia			
HealthLine	11.16	9.44	12.88
MayoClinic	12.88	11.26	14.51
MedicineNet	16.1	14.59	17.62
PsychCentral	13.48	11.9	15.07
WebMd	13.62	11.82	15.42
Wikipedia	10.32	8.47	12.17
Bulimia Nervosa			
HealthLine	11.82	9.9	13.74
MayoClinic	12.04	10.59	13.49
MedicineNet	15.02	13.55	16.49
PsychCentral	13.96	12.5	15.43
WebMd	14.98	13.6	16.36
Wikipedia	10.4	8.99	11.81
Panic Disorder			
HealthLine	11.54	10.17	12.91
MayoClinic	10.7	9.29	12.11
MedicineNet	12.52	11.3	13.75
PsychCentral	10.62	9.24	12
WebMd	12.3	11.07	13.53
Wikipedia	14.74	13.33	16.15
Substance abuse			
HealthLine	10.62	8.95	12.29
MayoClinic	11.12	9.72	12.52
MedicineNet	12.62	10.64	14.6
PsychCentral	12.96	10.61	15.31
WebMd	11.1	9.84	12.36
Wikipedia	14.2	12.75	15.66
Bipolar Disorder			

HealthLine	11.46	9.21	13.71
MayoClinic	15.24	13.38	17.1
MedicineNet	16.56	14.77	18.35
PsychCentral	13.9	12.33	15.47
WebMd	15.68	13.66	17.7
Wikipedia	11.66	10.01	13.31
PTSD			
HealthLine	11.48	9.87	13.09
MayoClinic	11.54	10.21	12.87
MedicineNet	14.64	13.32	15.96
PsychCentral	12.76	11.35	14.18
WebMd	15.44	13.99	16.89
Wikipedia	13.82	12.53	15.11
OCD			
HealthLine		9.16	12.84
MayoClinic	11.66	9.74	13.58
MedicineNet	14.9	13.4	16.39
PsychCentral	13.44	11.71	15.18
WebMd	13.78	12.25	15.31
Wikipedia	11.72	10.08	13.36
MDD			
HealthLine	14.8	13.06	16.53
MayoClinic	11.76	10.22	13.31
MedicineNet	14.18	12.6	15.76
PsychCentral	11.72	10.07	13.37
WebMd	14.6	12.88	16.32
Wikipedia	15.04	13.61	16.47
GAD			
HealthLine	12.6	11.31	13.89
MayoClinic	12.86	10.61	15.11
MedicineNet	16.04	13.92	18.16
PsychCentral	11.78	10.21	13.35
WebMd	12.4	10.64	14.16
Wikipedia	15.3	13.76	16.84
Social phobia			
HealthLine	11.8	9.86	13.74
MayoClinic	14.78	13.24	16.32
MedicineNet	15.1	13.04	17.16
PsychCentral	12.6	11.39	13.81
WebMd	14.44	12.87	16.01

Wikipedia	15.64	14.16	17.12
Schizophrenia			
HealthLine	13.02	11.15	14.89
MayoClinic	13.02	11.55	14.49
MedicineNet	16.36	14.56	18.16
PsychCentral	14.66	13.12	16.2
WebMd	13.54	11.83	15.25
Wikipedia	13.48	11.74	15.22

Discussion

The purpose of this paper was to provide a brief sampling of the analytic strategies available for analyzing nested reading grade-level data extracted from six different websites, for sixteen different mental health disorders and/or conditions, rated by five different readability indices. A discussion of various interpretations of the ICC was provided, as well as specific results from 1) a 2-level multi-level model with a random effect included to account for differences between raters/indices on average reading-level scores, and 2) a population-averaged GEE approach in which reading-level estimates were nested within a sample of all possible raters/indices.

In our example, data were conceptualized to be clustered within the various indices used to rate text extracted from online sources. Because we were not interested in exploring differences between raters, theorized that the readability indices selected were a random sample of all possible indices used to rate written text, and wished to retain data from all raters for each website and disorder combination in the modeling process, a marginalized models or GEE approach was selected as the best analytic strategy from a conceptual perspective. This approach was also selected given that the number of units of the grouping variable was small ($k = 5$ indices/raters), and some researchers suggest that utilizing multi-level modeling with a small number of groups may be inadvisable due to issues related to power and type I and II error (Hoyle & Gottfredson, 2015).

In this analysis, the variables website and disorder were treated as fixed effects, and an interaction term was included to account for differences in reading level scores across website and content area combinations. When comparing results

from the GEE approach and a 2-level multi-level model, it is apparent that these two strategies provided similar results, as expected. These differences are displayed below.

Table 3: Comparison of 95% CI obtained from 2-level MLM and GEE approaches

	Estimate	***MLM Lower 95% CL***	***MLM Upper 95% CL***	***GEE Lower 95% CL***	***GEE Upper 95% CL***
ADHD					
HealthLine	**14.36**	12.96	15.76	12.43	16.29
MayoClinic	**12.36**	10.71	14.01	10.43	14.29
MedicineNet	**15.56**	13.61	17.52	13.63	17.49
PsychCentral	**13.56**	12.44	14.68	11.63	15.49
WebMd	**8.9**	7.21	10.6	6.97	10.83
Wikipedia	**13.08**	11.61	14.55	11.12	15.01
Agoraphobia					
HealthLine	**11.16**	8.85	13.47	9.23	13.09
MayoClinic	**12.14**	10.97	13.31	10.21	14.07
MedicineNet	**16.54**	14.93	18.15	14.61	18.47
PsychCentral	**12.42**	10.81	14.03	8.49	16.45
WebMd	**5.62**	4.17	7.07	3.69	7.55
Wikipedia	**13.5**	12.23	14.77	11.57	15.43
Alcoholism					
HealthLine	**10.98**	9.23	12.73	9.05	12.91
MayoClinic	**10.18**	9.11	11.25	8.25	12.11
MedicineNet	**16.46**	14.86	18.06	14.53	18.39
PsychCentral	**12.46**	10.93	13.99	10.53	14.39
WebMd	**13.42**	11.91	14.93	11.49	15.35
Wikipedia	**11.2**	9.83	12.57	9.27	13.13
Anorexia Nervosa					
HealthLine	**11.58**	9.67	13.49	9.65	13.51
MayoClinic	**12.64**	11.26	14.02	10.71	14.57
MedicineNet	**15.32**	13.88	16.76	13.39	17.25
PsychCentral	**13**	11.45	14.55	11.07	14.93
WebMd	**14.1**	12.66	15.54	12.17	16.03
Wikipedia	**10.12**	8.62	11.62	8.19	12.05
Bipolar Disorder					
HealthLine	**11.46**	9.21	13.71	9.53	13.39
MayoClinic	**15.24**	13.38	17.1	13.31	17.17
MedicineNet	**16.56**	14.77	18.35	14.63	18.49
PsychCentral	**13.9**	12.33	15.47	11.97	15.83

WebMd	**15.68**	13.66	17.7	13.75	17.61
Wikipedia	**11.66**	10.01	13.31	9.73	13.59
Borderline PD					
HealthLine	**12.38**	10.54	14.22	10.44	14.31
MayoClinic	**11.58**	10.11	13.05	9.64	13.51
MedicineNet	**17.9**	16.09	19.71	15.97	19.83
PsychCentral	**17.36**	15.38	19.34	15.43	19.29
WebMd	**9.32**	7.58	11.06	7.39	11.25
Wikipedia	**16.4**	14.67	18.33	14.47	18.33
Bulimia Nervosa					
HealthLine	**11.82**	9.9	13.74	9.89	13.75
MayoClinic	**12.04**	10.59	13.49	10.11	13.97
MedicineNet	**15.02**	13.55	16.49	13.09	16.95
PsychCentral	**13.96**	12.5	15.43	12.03	15.89
WebMd	**14.98**	13.6	16.36	13.05	16.91
Wikipedia	**10.4**	8.99	11.81	8.47	12.33
GAD					
HealthLine	**12.6**	11.31	13.89	10.67	14.53
MayoClinic	**12.86**	10.61	15.11	10.93	14.79
MedicineNet	**16.04**	13.92	18.16	14.11	17.97
PsychCentral	**11.78**	10.21	13.35	9.85	13.71
WebMd	**12.4**	10.64	14.16	10.47	14.33
Wikipedia	**15.3**	13.76	16.84	13.37	17.23
MDD					
HealthLine	**14.8**	13.06	16.53	12.87	16.73
MayoClinic	**11.76**	10.22	13.31	9.83	13.69
MedicineNet	**14.18**	12.6	15.76	12.25	16.11
PsychCentral	**11.72**	10.07	13.37	9.79	13.65
WebMd	**14.6**	12.88	16.32	12.67	16.53
Wikipedia	**15.04**	13.61	16.47	13.11	16.97
OCD					
HealthLine	**11**	9.16	12.84	9.07	12.93
MayoClinic	**11.66**	9.74	13.58	9.73	13.59
MedicineNet	**14.9**	13.4	16.39	12.97	16.83
PsychCentral	**13.44**	11.71	15.18	11.51	15.37
WebMd	**13.78**	12.25	15.31	9.85	13.71
Wikipedia	**11.72**	10.08	13.36	9.79	13.65
Panic Disorder					
HealthLine	**11.54**	10.17	12.91	9.61	13.47
MayoClinic	**10.7**	9.29	12.11	8.77	12.63

MedicineNet	**12.52**	11.3	13.75	10.59	14.45
PsychCentral	**10.62**	9.24	12	8.69	12.55
WebMd	**12.3**	11.07	13.53	10.37	14.23
Wikipedia	**14.74**	13.33	16.15	12.81	16.67
PTSD					
HealthLine	**11.48**	9.87	13.09	9.55	13.41
MayoClinic	**11.54**	10.21	12.87	9.61	13.47
MedicineNet	**14.64**	13.32	15.96	12.71	16.57
PsychCentral	**12.76**	11.35	14.18	10.83	14.69
WebMd	**15.44**	13.99	16.89	13.51	17.37
Wikipedia	**13.82**	12.53	15.11	11.89	15.75
Schizophrenia					
HealthLine	**13.02**	11.15	14.89	11.09	14.95
MayoClinic	**13.02**	11.55	14.49	11.09	14.95
MedicineNet	**16.36**	14.56	18.16	14.43	18.29
PsychCentral	**14.66**	13.12	16.2	12.73	16.59
WebMd	**13.54**	11.83	15.25	11.61	15.47
Wikipedia	**13.48**	11.74	15.22	11.55	15.41
Social phobia					
HealthLine	**11.8**	9.86	13.74	9.87	13.73
MayoClinic	**14.78**	13.24	16.32	12.85	16.71
MedicineNet	**15.1**	13.04	17.16	13.17	17.03
PsychCentral	**12.6**	11.39	13.81	10.67	14.53
WebMd	**14.44**	12.87	16.01	12.51	16.37
Wikipedia	**15.64**	14.16	17.12	13.71	17.57
Specific Phobia					
HealthLine	**11.16**	9.44	12.88	9.23	13.09
MayoClinic	**12.88**	11.26	14.51	10.95	14.81
MedicineNet	**16.1**	14.59	17.62	14.17	18.03
PsychCentral	**13.48**	11.9	15.07	11.55	15.41
WebMd	**13.62**	11.82	15.42	11.69	15.55
Wikipedia	**10.32**	8.47	12.17	8.39	12.25
Substance abuse					
HealthLine	**10.62**	8.95	12.29	8.69	12.55
MayoClinic	**11.12**	9.72	12.52	9.19	13.05
MedicineNet	**12.62**	10.64	14.6	10.69	14.55
PsychCentral	**12.96**	10.61	15.31	11.03	14.89
WebMd	**11.1**	9.84	12.36	9.17	13.03
Wikipedia	**14.2**	12.75	15.66	12.27	16.13

The similarities in results generated from both modeling strategies may be due to a number of factors. These factors may include the continuous nature of the outcome variable, as well as the limited number of factors included as fixed effects in the model. Although there are some differences in the interpretation of outcomes between multi-level models and GEE when the outcome variable is binary or non-linear, the interpretation is largely consistent across models for continuous data. Likewise, although the number of groups included to account for clustering within the data was small (scores nested within five raters), only the variables website and disorder were included as explanatory variables in both models. In this case, the decision to utilize GEE over a 2-level multi-level model hence lies in the fundamental question of interest to the researcher. Given that the primary research objective of this study was to evaluate the relationship between websites, disorders, and their interaction on reading grade-level scores across a population of possible raters (indices), a GEE or marginal models approach was hypothesized to be the best conceptual fit for this specific question. However, because the outcome data are linear and normally distributed, multi-level modeling may also be an appropriate alternative strategy.

Overall, aside from a key few instances, the reading grade level for all disorders across the various websites explored far exceeded the suggested 6^{th} to 8^{th} grade reading level guidelines established by the CDC and other similar organizations. In some cases, (i.e. text related to borderline personality disorder from MedicineNet.com), the estimated reading grade level reached as high as 17.9. This estimate suggests that, on average, an individual with an advanced graduate degree

(grade 17.9) would be able to read the selected text effectively. In other instances, (i.e. text related to ADHD and Agoraphobia from WebMd.com), reading grade level estimates were much lower, and consistent with a 6th to 8th grade reading level, respectively. These estimates suggest that an individual who completed the 6th to 8th grade could effectively read the selected text. However, all other estimates obtained were markedly higher, with a minimal average high school reading level required to adequately read the selected text.

Interestingly, text related to borderline personality disorder demonstrated the highest reading grade level estimate, followed by text related to bipolar disorder, social phobia, schizophrenia, MDD, and GAD, in descending order of grade level. Examination of estimates for these disorders generally suggests that an individual with an average post-high school reading level could effectively read the segments of text selected for analysis. Given the severity of impairment often associated with these disorders (particularly borderline personality disorder, bipolar disorder, and schizophrenia), it could be surmised that the information available online from the websites surveyed is not only relatively inaccessible to most healthy consumers, but also especially to those struggling with serious mental illness.

Not surprisingly, little difference was noted in reading grade level estimates between MDD and bipolar disorder, as these disorders may share a common language regarding general symptoms of depression. Likewise, given similarities in language, symptom presentation, and etiology, there was no notable difference in reading level scores for alcoholism and substance abuse, as well as social phobia and specific phobia. However, this rationale could not be extended to text describing the two

predominant eating disorders examined in this study: reading level estimates for bulimia nervosa were significantly higher than those for anorexia nervosa. It is possible that further exploration of text *content* may reveal emphasis on different features, symptoms, or etiology of each disorder, hence contributing to differences in reading level estimates.

Future research may focus on: 1) increasing the number of clusters of the grouping variable by including ratings from additional indices; 2) re-conceptualizing the data as being nested within various websites, or within disorders (instead of within raters) to expand the number of groups; 3) further investigating inter-rater reliability by asking multiple individuals to extract text from the websites selected for the study; 4) investigating how the construct of reading comprehension is related to the readability of selected text using human subjects; and 5) exploring how readability and comprehension are related to utilization of health services. These ideas for future investigation may address some of the key limitations of this study, which include a small number of groups of the clustering variable, and the absence of any information regarding how reading comprehension might be related to reading-grade level of selected text. Furthermore, only information from disorders that were available on all web platforms was selected for this analysis. Expanding the number of websites and disorders for analysis may provide a more comprehensive picture of the readability of online mental health materials, and may reveal additional or alternative associations not demonstrated in this analysis.

Overall, despite some differences in the width of confidence intervals, results from the multi-level modeling and GEE approach are consistent in that they suggest

that although some website and disorder combinations had higher readability scores than others, scores from all websites and for all disorders exceeded the recommended 6^{th} to 8^{th} grade standard. This result is important because it demonstrates that much of the material obtained online is not written at a level that is comprehensible for the majority of consumers in the United States. In order to prevent the perpetuation of existing health disparities associated with lack of health literacy, writers of public online mental health materials are advised to take great care in ensuring that the information they post is accessible to as many individuals as possible. Readers are also encouraged to explore alternative modeling strategies for more complicated data, depending on their primary research aim.

References

Baker, L., Wagner, T.H., Singer, S., & Bundorf, K. (2003). Use of the internet and email for health care information. *JAMA, 289*(18), 2400-2406.

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions: A comparison of intraclass correlation coefficients, rwg (j), hierarchical linear modeling, within-and between-analysis, and random group resampling. *The Leadership Quarterly*, *13*(1), 69-93.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529-569.

Diaz, J.A., Griffith, R.A., Ng, J.J., Reinert, S.E., Friedmann, P.D., & Moulton, A.W. (2002). Patients' use of the internet for medical information. *Journal of General Internal Medicine, 17*(3), 180-185.

Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health*, *21*(1), 171-192.

DuBay, W. H. (2004). *The principles of readability*. Online Submission.

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass*, *3*(6), 979-991.

Eysenbach, G., & Kohler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ, 324*,573-576.

Hamilton, L. (2012). *Statistics with Stata: version 12*. Cengage Learning.

Hayes, A.F. (2006). A primer on multilevel modeling. *Human Communication Research, 32,* 385 – 410.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. New York, NY: Routledge.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, *23*(6), 723-744.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*(2), 157-174.

Hoyle, R. H., & Gottfredson, N. C. (2015). Sample size considerations in prevention research applications of multilevel modeling and structural equation modeling. *Prevention Science*, *16,* 987-996.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*(4), 467-474.

IBM Corp. Released 2012. *IBM SPSS Statistics for Windows*, Version 21.0. Armonk, NY: IBM Corp.

Kaplan, D. (Ed.). (2004). *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology (In D. Gilbert, S. Fiske, & G. Lindzey (Eds.). *The handbook of social psychology* (Vol. 1, pp. 233–265).

Kutner, M., Greenburg, E., Jin, Y., & Paulsen, C. (2006). The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. *National Center for Education Statistics*.

Luke, D. A. (2004). *Multilevel modeling* (Vol. 143). Thousand Oaks, CA: Sage.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86-92.

McInnes, N., & Haglund, B. J. (2011). Readability of online health information: implications for health literacy. *Informatics for Health and Social Care*, *36*(4), 173-189.

McMullan, M. (2006). Patients using the Internet to obtain health information: How this affects the patient-health professional relationship. *Patient Education and Counseling*, *63*, 24-28.

Neuhauser, L., & Paul, K. (2011). Readablity, comprehension, and usability. In B. Fischhoff, N. T. Brewer & J. S. Downs (Eds.), *Communicating risks and benefits: An evidence-based user's guide* (129-148). New Hampshire: US Department of Health and Human Services – Food and Drug Association.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85-112.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167-190.

Raudenbush, S. W. (2000). [Marginalized Multilevel Models and Likelihood Inference]: Comment. *Statistical Science, 15,* 22-24.

Raudenbush, S. W. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Newbury Park, CA: Sage.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420 – 428.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*, 31-65.

Warne, R. T., Li, Y., McKyer, E. L. J., Condie, R., Diep, C. S., & Murano, P. S. (2012). Managing clustered data using hierarchical linear modeling. *Journal of Nutrition Education and Behavior, 44*(3), 271-277.

Wears, R.L. (2008). Advanced statistics: Statistical methods for analyzing cluster and cluster-randomized data. *Academic Emergency Medicine, 9*(4), 330 – 341.

Weiss, B. D. (2003). Health literacy. *A manual for clinicians. Chicago: American Medical Association Foundation and American Medical Association*.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics, 44,* 1049-1060.