The wQUADAS: Creation and reliability

Rachael Kang

University of North Carolina, Chapel Hill

Abstract

Within this study we attempted to create an objective assessment that would measure website quality to dispel the stereotype that information found on Wikipedia is inaccurate, ultimately to encourage clinicians to utilize the evidence based assessments (EBA) pages on Wikipedia. Over 900 participants were recruited from Amazon's Mechanical Turk (MTurk) and were tasked to rate one webpage using the assessment and one webpage not using the assessment as either quality or not. An exploratory factor analysis revealed that our assessments measured four factors of website quality: style, maintenance/coverage, coverage, and authority. Correlational data supported style as one of the four factors, and suggests that style is an important factor to those rating the quality of a webpage. At the end of the study, over 70% of participants indicated that they would use the measure again, showing promise that this measure may make a difference in how websites such as Wikipedia are perceived and utilized.

Keywords: webpages, Wikipedia, Evidence based Assessments, wQUADAS

The wQUADAS: Creation and reliability

Evidence Based Assessments

The dissemination and utilization of evidence-based assessments (EBA) has faced an unusual amount of resistance in the field of psychology. EBA are assessments that are grounded in research and have data that support the reliability and validity of the assessment. Ideally, EBAs would be used in clinics around the world to ensure that all clinicians were using the same caliber of assessment to diagnose and assess mental illnesses in their clients. By doing so, the rates of misdiagnoses and delayed diagnoses would decrease, bettering patient outcomes. One explanation clinicians may not use EBA could be a lack of clear communication between the researchers creating the new assessments and clinicians. Researchers are constantly finding new data and creating new assessments for mental illnesses, but they are not communicating their findings to clinicians in a clear and effective manner. Another reason is that clinicians will typically only use unstructured or semi-structured interviews when assessing patients, rather than using a battery of assessments as normally recommended (Jenson-Doss & Hawley, 2010). Finally, and possibly the biggest dilemma clinicians face in using EBA, is that it is difficult for clinicians everywhere to always be knowledgeable about which assessments have the strongest psychometric properties, are the most applicable to their clinical population, and are the most up-to-date with current diagnostic criteria, while also managing their clinical loads. All these scenarios make it difficult for clinicians to easily implement EBA in their practices.

Addressing this last concern, if there were a place that housed all EBAs, including their psychometric properties and histories of creation, that was easily accessible and navigable, perhaps clinicians would be more inclined to use EBAs. Unfortunately, though there have been attempts, no site or journal has been able to create and maintain an accessible and navigable database of these most updated, psychometrically strong, “gold-standard” assessments. The Society of Child Clinical and Adolescent Psychology (SCCAP) made initial efforts in mid-2010 to collect and disseminate evidence based assessments through a website named effectivechildtherapy.org. However, the difficulty of constantly maintaining an up-to-date site, the cost of updating the site, as well as low traffic to this page in comparison to other pages on the Internet made it clear that this was not the best platform of dissemination. Dr. Eric Youngstrom and his graduate student Mian Li Ong (M.A.) explored other web platforms such as blogs and wordpress websites, but the problem of low traffic as well as difficulty finding a way to constantly update the site forced them to break the mold and look toward Wikipedia.org.

Wikipedia as a Platform of Dissemination

Wikipedia.org is a promising platform for disseminating EBA. Wikipedia is a free, open, online encyclopedia that allows for the mass distribution of up-to-date information to a huge number of people. As of January 31, 2018, according to Alexa Internet, Wikipedia was ranked the fifth most visited site on the Internet. Further, Wikipedia and Google have an agreement whereby as long as Wikipedia keeps its information free, if there is a Wikipedia page about a Google-searched topic, the Google search engine will display that Wikipedia entry on the first page of search results. All these features give Wikipedia the potential to revolutionize the dissemination of evidence-based assessments.

Wikipedia offers a solution to most EBA dissemination barriers as previously mentioned. Because of Wikipedia’s partnership with Google, all clinicians would need to do is perform a Google search on an assessment. They will be given the link to a Wikipedia page about that assessment, and, if the assessment is not copyrighted, a pdf of the assessment can be attached to the Wikipedia page as well. Further, all EBA measures can be tagged on Wikipedia as “EBA”, creating the possibility for the centralization of all EBA pages on Wikipedia. This solves the problem of lack-of-access to EBA material and the lack of one, centralized database of EBA pages. Additionally, clinicians would easily be able to Google search any number of assessments, and effortlessly peruse the information about them on the Wikipedia pages, and compare assessments to each other to find the best assessment for their practice. All these potential uses of Wikipedia could change the dissemination of EBA for the better.

Furthermore, recent research supports the accuracy of information found on Wikipedia. In a study done by Flanagin & Metzger (2011), researchers discovered that information found on Wikipedia articles about a certain topic was just as accurate, if not more accurate than corresponding information found in *Encyclopedia Britannica*. Specifically relating to psychology, another study found that information on Wikipedia regarding depression and schizophrenia was rated as "higher quality" by mental health professionals and by depression and schizophrenia experts than printed information (e.g., a textbook; Reavley, Mackinnon, Morgan, & Alvarez-Jimenez, 2011). On more heavily visited pages, such as the Major Depression page, Wikipedia has added features that require a certain amount of Wikipedia editing experience before being allowed to make an edit on that page. Wikipedia's open source nature is what ensures that information on Wikipedia pages is kept up-to-date and accurate. Though studies are sparse, the existing studies suggest that Wikipedia may be a good platform where credible and reliable information can be placed and does not become lost.

Despite these positive indications of Wikipedia being a good place for the dissemination of EBAs, the perception that Wikipedia is not a reliable source of information may leave clinicians viewing Wikipedia as a source of information in a negative light, stopping them from utilizing Wikipedia pages. Ajzen's Theory of Planned Behavior (TPB) posits that an individual's behavior is a function of the intention to perform that behavior (Ajzen, 1991). Meta-analyses of TPB propose that an individual's attitude toward a behavior is a strong indicator of the intention to perform that behavior (Armitage & Conner, 2001; Godin & Kok, 1996). As such, it is reasonable to assume that if a clinician's negative view of looking up information on Wikipedia stops the clinician from using Wikipedia, changing an individual's attitude from negative to positive will increase the intention to look to Wikipedia for information, thus increasing the behavior of an individual looking to Wikipedia for information.

In the present study, I attempt to create an assessment that acts as an objective measure of the quality of a webpage. By using such a measure, a more objective quality rating of Wikipedia could be obtained that could change people's opinions that Wikipedia is not a reliable source of information. If successful, Wikipedia could be used to help proliferate the use of EBA in every day practice.

The present study: the creation and reliability of the wQUADAS

This study aims to create an assessment tool, the website version of the quality assessment of diagnostic accuracy studies (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003) or the wQUADAS, that will objectively measure the quality of websites and of Wikipedia psychological assessment pages. Creating an assessment tool that objectively measures the quality of not only websites, but also Wikipedia pages, will help combat the perception that the Internet is an unreliable source of information. The wQUADAS will ultimately aid in the dissemination of evidence-based assessments not only to clinicians, but also to the general public. There are three research questions addressed in the study:

Question 1: How many factors of webpage quality will the wQUADAS measure?

Question 2: How will people's ratings of websites change after taking the wQUADAS?

Question 3: How will people's opinions of the quality of a webpage correlate with their perception of the page at first glance?

Method

Creating the wQUADAS

This study aims to create and establish the reliability of the website version of the Quality Assessment of Diagnosis Accuracy Studies (Whiting et al., 2003), the wQUADAS. An initial pool of 30 items was created and evaluated by a focus group that included the principle investigator (Rachael Kang, BS), faculty advisor (Eric Youngstrom, Ph.D), and graduate student advisor (Mian Li Ong, MA). Each member of the focus group completed the wQUADAS to rate a website (www.mentalhealthamerica.net/conditions/bipolar-disorder), and scores were compared to one another. The focus group talked through each question as one would do in a cognitive interview to understand how each member was interpreting the question and how to answer it. After the cognitive interviews, focus group members' scores on the wQUADAS for the mentalhealthamerica.net website better matched one another. One more website was rated (www.add.org/adhd-facts) by the same people in the focus group, and two other members of the Mood Emotions Clinical Child Adolescent Assessment (MECCA) lab, a lab led by Dr. Eric Youngstrom. These ratings were consistent with one another, indicating to the principle investigator that data collection for the items was ready. The final list of wQUADAS items can be seen in Figure 4.

Participants and procedures

Participant recruitment occurred via Amazon's Mechanical Turk (MTurk). MTurk is a new tool utilized by researchers to gather large quantities of data at lower costs and more efficient time rates through crowdsourcing (Buhrmester, Kwang, & Gosling, 2011). MTurk, owned by Amazon, allows users to give surveys to others who are paid to complete surveys. In this study MTurk users were paid $0.20 a survey in order to maximize the *N* to follow the rule of 500 given by Comrey and Lee (1992), which suggests a sample size of 500 as a “very good” size for factor analyses, but a sample size of 1000 being “excellent.” Sample size also depends on the strength of the indicators, and also the number of items per factor. Inclusionary criteria included individuals 18 years or older who were able to read and understand English at a conversational level, and were located within the US. There were no exclusionary criteria.

Figure 1 provides a flow chart about which websites participants were assigned to rate in addition to the websites that were chosen for the study. As seen in Figure 1, in the box on the far left, it states that MTurk randomly provided one of six survey links to the participants, which led participants to one of six surveys titled “wQUADAS survey X part 1”, which will be referred to as survey part 1 from herein. At the start of the survey, participants answered non-identifying demographic questions about highest level of education attained at time of survey, race, and sex. Once participants provided demographic information, they were asked to rate the quality of one of the six webpages listed on Figure 1. For example, if a participant was given the link to wQUADAS survey 1 part 1, he or she was asked to rate the quality of a depression blog on www.prevention.com on a scale of 0 to 3 based on his or her opinion. At the completion of part 1, as seen in Figure 1 again, participants would be automatically directed to a pre-assigned part 2 survey where they would rate another website, only this time they would use the wQUADAS. This second survey will herein be referred to as survey part 2. For example, if a participant was assigned wQUADAS survey 1 part 1, at the end of that survey they would automatically be asked to rate wQUADAS survey 5 part 2, a www.webmd.com page about suicidal ideation. The full list of wQUADAS items can be seen in Figure 3.

As inspired by Cummings (2009), the pages in this study were chosen based on their popularity within topics about mental illness on Google trends during 2017 over the last year. The following topics were the most Googled terms related to mental illnesses: depression, anxiety, and suicide. Google-advertised pages (Google search results marked with “ad”) were ignored, as they were mostly pages advertising a treatment for depression, anxiety, or suicidal ideation. Further, government sponsored webpages were not selected due to the assumption that government sponsored webpages (pages ending in .gov) would create too much of a bias in favor of those webpages, and would inappropriately skew quality ratings.

Measures

Websites-Quality Assessment of Diagnosis Accuracy Studies (wQUADAS). The wQUADAS is an assessment of the quality of a webpage. The wQUADAS contains 30 items. Items are answered on a 4-point Likert-type scale (e.g., 0-not easy, 1-somewhat easy, 2-easy, 3-very easy). The content is designed to reflect standards of reliable secondary sources of information (Metzger, 2007; Metzger & Flanagin, 2013). Items are broken down into five factors, which refer to the five criteria of a reliable secondary source as described by Metzger (2007): objectivity, authority, accuracy, coverage, and currency. Figure 3 displays the grouping of the items. Summing up all item responses provided the score of the wQUADAS (Min = 0, Max = 90).

Data analytic plan

Preliminary analysis. Each participant provided two quality assessments on two out of six potential different webpages, once using the wQUADAS and once without using the wQUADAS. Any participant who spent less than a minute completing the assessment, or over one hour, was excluded from the data. After excluding incomplete data, descriptive and frequency statistics determined the demographic make-up of the study population. Next, the data were split to compare race, sex, and education levels between wQUADAS ratings of "Not quality" and "Quality" for each of the six webpages. A Chi-squared statistic was used to test the differences in gender, race, and education between the "Not quality" and "Quality" ratings, and *phi* was calculated to determine effect size. Independent groups *t*-tests measured the differences in the mean ages and scores on the wQUADAS between the "Not quality" and "Quality" ratings. *Phi* was calculated to determine effect size for gender, race, and education, and Cohen's *d* was used to determine effect size of the wQUADAS rating. Cohen's *d*, a common way of calculating effect size, can be used to support utility (Larner, 2014). Further, Cohen's *d* can be used to test the "power" of a *t*-test based on the population size in order to reduce Type I or II errors (Cohen, 1992). All statistical were conducted using SPSS (Version 24.0).

Exploratory factor analysis. A factor analysis is a statistical model that expresses random variables, in this case items on the wQUADAS, as linear functions of common factors (Jollife, 2002). Velicer's Minimum Average Partial Test (Velicer, 1976) and Parallel analysis were both used to calculate the number of factors being measured by the wQUADAS as well as the loadings for each item. Factor loadings describe which items best load, or "match", with each factor. By determining the number of factors and which items loaded on, or belonged to, which factor, the content validity of the wQUADAS can be better established. When conducting the factor analysis, an oblique, Promax, rotation was used because it was assumed one or more factors would correlate with one another.

Paired samples *t*-test. A paired samples *t*-test measures the difference in means of two groups that are paired in some way. In this study, the samples were paired by the survey part 1 and survey part 2 groupings as the survey part 1 group did **not** use the wQUADAS and the survey part 2 group **did** use the wQUADAS. Paired sample *t*-tests will determine if there is a significant difference in the means of quality ratings in the survey part 1 and survey part 2, as well as if there is a significant difference in the confidence of quality assessment in the survey part 1 and survey part 2 groups.

Inter-rater and internal reliability. Reliability is an important feature of a measure as a measure cannot be valid if it is not first reliable (Callans, 2012). Internal consistency refers to how well items correlate with one another. For the wQUADAS, Cronbach's α was calculated to determine the internal consistency of the items in the measure as a whole. More important than internal consistency, however, is the inter-rater reliability of the wQUADAS. Because multiple raters measured the quality of the same webpages using the wQUADAS, inter-rater reliability, or the consistency between two or more raters, is a key element in assessing the quality of a measure. This is measured by Fleiss's κ . Fleiss's κ was preferred over Cohen's κ because more than 3 raters were being compared to one another to assess inter-rater reliability (McHugh, 2012).

Regression. Hierarchical logistic and linear regressions tested to see how demographic variables related to the score on the wQUADAS and the quality rating given by the participant. For the linear regression, three models were calculated to see how race, gender, and education levels predicted the scores on the wQUADAS. Model 1 entered race, model 2 entered gender, and model 3 entered education level. The logistic regression had the same model, but tested the effect of the demographic variables on the quality rating given by the participants. Quality ratings were determined by taking the participant's rating on the 0 to 3 scale and recoding the values such that scores of 0 and 1 were "Not quality" and scores of 2 and 3 were considered "Quality". The linear regression's R^2, logistic regression's Nagelkerke R^2 ratios, and both log-likelihood ratios were reported to describe the fit of the regression lines as well as the significance.

Results

Descriptive Statistics

In survey part 1 of the study, an $N = 1010$ (59% Female, 70% White) was collected from MTurk with no participant data being excluded. In survey part 2 of the study, an original N = 1088 was collected from MTurk, but a final $N = 931$ (56% Female, 74% White) was attained when participants who took over 3600 seconds or under 60 seconds (the lowest and highest intervals of time) were excluded from the analyses.

Table 1 displays the results for a chi-squared statistic that was used to test differences between quality and not quality ratings, both using and not using the wQUADAS, for all 6 websites among the variables gender, education, and race. A quality rating was attained by having participants rate the website on a scale of 0 to 3 to determine page quality, then recoding the ratings of 0 or 1 as "Not quality" and recoding ratings of 2 or 3 as "Quality". In the survey 1 group, or the non-wQUADAS group, race had a significant difference between ratings of quality and not quality ($X^2(4) = 11.04$, $p = 0.03$), while in the survey 2 group, or the wQUADAS group, gender had a significant difference between ratings of quality and not quality ($X^2(1) = 4.58$, $p = 0.03$). However, both of these results had small effect sizes (*phi* < 0.2).

Tables 2 to 7 also shows the results of chi-squares, but they focus on each individual website and if any one particular website has a large difference in quality vs. not quality ratings among sex, race, and education level. As seen in Table 5, education had a small significant difference in giving a good quality rating to the Wikipedia page on the in the non-wQUADAS ($p < .05$) for the health.com page about Anxiety. Table 7 showed a small significant difference between in education in the wQUADAS quality rating (p < .05) for the Columbia Suicide Severity Rating Scale (CSSRS).

The logistic and linear regressions results are displayed in Table 8. Once again, it appears that there is a small potential that gender may act as a covariate of wQUADAS score and quality rating. However, both the linear regression and logistic regression models only accounted for 1% of the variance seen (Negelkerke R^2 = 0.01; R^2 = 0.01).

The wQUADAS had a Cronbach's α = 0.90 and a Fleiss' κ = 0.11, indicating strong internal consistency but mediocre inter-rater reliability. The internal consistency, Cronbach's α, did not change much with any one item being deleted, indicating all items correlated well with each other.

wQUADAS Exploratory Factor Analysis

The wQUADAS score was obtained by summing all items. Table 9 shows the mean wQUADAS score for each website (Minimum score = 0, Maximum score = 90). The Wikipedia depression page scored the highest on average, with the depression prevention.com page coming in second, followed by Webmd's suicide page.

Figure 2 shows a scree plot of potential factors. There seems to be six distinct factors with factors 5 and 6 having similar Eigen values. However, after conducting MAP and parallel analyses, only 4 factors had four or more item loadings. An oblique, Promax, rotation was used as it was expected the factors would correlated with one another. Table 12 reports the factor correlations. Factors 1 and 4 correlated the least with one another while factors 2 and 3 correlated the highest with one another. Figure 3 shows what the ideal factor loading would have been (e.g., factor 1: objectivity, with items 1-6 loading onto it), and Figure 4 shows the results of the actual factor loadings.

Factor 1 was originally to be objectivity, but the items that loaded on factor 1 more reflects first impressions of the webpage (e.g., grammatical errors). Factor 2 was meant to be authority, but now looks to be a blend of both website maintenance and coverage of sources. Factor 3 was initially accuracy, but looks more like coverage of topic. Finally, Factor 4 was coverage, but is now clearly authority.

Pre-post wQUADAS analyses

A paired samples *t*-test of a pre-wQUADAS rating and a post-wQUADAS rating showed a significant decrease ($p < .0001$) in quality rating across all 6 websites after having taken the wQUADAS (Table 9). Table 10 shows the results of another paired samples *t*-test that indicates that the confidence of a person's quality rating in the pre-wQUADAS and post-wQUADAS across all the webpages decreased. However, the only significant changes in this t-test were seen for the Wikipedia depression page ($t(137) = 3.68$, p < .001), Health.com anxiety page ($t(113) = 2.17$, $p < .05$), and the WebMD suicidal ideation page ($t(111) = 3.61$, $p < .001$).

Table 11 shows the results of a Pearson's correlation that indicates a high and significant correlation between a person's pre-read impression (or first impression) of the quality of a webpage and a person's quality rating of a webpage after reading the page (or post-read impression) ($r(1086) = 0.62$, $p < .0001$).

Discussion

The purpose of this study was to create an objective measure of website quality so that it can be used to change people's opinions about the credibility of information found on Wikipedia, ultimately to allow for the dissemination of EBA through Wikipedia. The first question asked in this study is how many factors would be measured by the wQUADAS. Ideally, we expected to see five factors measured by the wQUADAS because the items were created and grouped based on Metzger's (2007) article on what makes a good Internet source. However, after running both a MAP and parallel exploratory factory analysis, only four factors were measured. The four factors were renamed as style, freshness, coverage, and authority. Freshness describes the combination of both maintenance and currency. The addition of style as a factor indicates that the layperson may use the aesthetic of a website (e.g., the professionalism, how grammar free, general page layout, etc…) to judge the quality of the page.

Correlational data support this supposition as people's first impressions of a page highly correlated with their post-read impressions of that same page ($r(1008) = 0.62$, $p < .001$), suggesting that even though a webpage may not actually be a good quality site, if it is aesthetically pleasing, a person may still rate the quality of the page more favorably than it actually is. This finding begs another question of potential research: how heavily do people rely on the aesthetics of a page to determine its quality?

This correlational data also answers the third question of this study that asks about the relationship between a participant's first impression of the webpage and their post-read impression of the same page. There is a high, positive correlation between someone's rating based off their first impression of the page and their post-read rating of the page. Furthermore, when looking at the pages that did get high ratings on the wQUADAS as well as high ratings without the wQUADAS, it was noted that these highly rated sites all had interactive features as well as sleek and easy-to-look-at web designs that allow the reader to comfortably peruse through information.

The second question of this study asked if people's quality ratings of webpages would change after taking the wQUADAS. The paired samples *t*-test showed quality ratings were significantly lower in the post-wQUADAS group than in the pre-wQUADAS group (Table 9).

Additionally, the confidence of participant ratings dropped between the pre-wQUADAS and post-wQUADAS, but only the confidence ratings for the Wikipedia depression page, Health.com anxiety page, and WebMD suicide page had a significant decrease (Table 10). One possible explanation is that the introduction of the wQUADAS gave participants a rubric on how harshly to rate the websites, versus letting participants rate the websites however they saw fit. However, because part 2 of the survey was not randomly assigned (participants were automatically redirected from the part 1 website to a pre-determined part 2 website), we are unable to determine whether the changes seen in both the quality ratings and the confidence ratings are due to the introduction of the wQUADAS or some other moderating variables. A testing effect may have had significant influence on the participant's confidence and quality rating of the second website due to the non-randomization of the post-wQUADAS group. This acts as one of the limitations of this study.

Another limitation is the potential effect of negative salience incurred when using the wQUADAS to rate the quality of webpages. Negative salience is the effect seen where originally positive impressions are changed to negative impressions when there is more negative information presented than positive information (Richey, Mcclelland, & Shimkunas, 1967). Further, according to Richey et al., (1967), an originally negative impression is not changed to positive impression even if there is more positive information presented to outweigh the negative information. In this study, participants' answers on the wQUADAS would have acted as the source of information on the quality of the webpage. If that source of information seemed to be more negative, participant's impressions of the page may have changed to negative or been reinforced as negative, causing the participant to rate the page as low quality, or become increasingly stricter when answering the remaining wQUADAS items. Thus leading the participants to rate the pages as poorer quality than actually are.

Despite these limitations, data from the study are able to give a better understanding of what a person considers important when rating the quality of a webpage. For example, it appears that the appearance of a website is just as influential as the actual content of the page on a person's quality rating of the website. Future research should explore any other potential latent factors influencing raters' evaluation of webpage quality, as well as further tease apart the factors being measured in factor 2 of the wQUADAS.

Additionally, other avenues of future research could look at the relationship between page aesthetic, page content, and page quality, to see whether page aesthetic or page quality plays a larger role in determining page quality. There could also be room for a validity study to see if the wQUADAS is accurately measuring quality websites as quality.

Conclusion

This study was a preliminary study on the construction of the wQUADAS, an objective measure of webpage quality. Having a measure such as this could change the way people perceive open-platform sites such as Wikipedia, which are currently viewed as unreliable sources of information. By changing this view to something more positive, researchers working on developing EBAs could utilize Wikipedia to make EBAs more accessible to a wider audience. 71% of wQUADAS respondents indicated they would use the wQUADAS again to rate the quality of websites. This response rate suggests that people would likely use the wQUADAS again to rate the quality of the website, alluding to the potential popularity of using the wQUADAS in rating the quality of a website.

Furthermore, having a measure such as this could also change the way web-designers and researchers put up information on the Internet. It would give them a guideline as to what to include on their respective webpages to make their sites look more credible and reliable compared to other sites on the Internet. Both reasons discussed above will contribute to the wider dissemination of EBA to clinics across the world**.**

References

Ajzen, I. (2002). Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior1. *Journal of Applied Social Psychology,32*(4), 665-683. doi:10.1111/j.1559-1816.2002.tb00236.x

Armitage CJ, Conner M. Efficacy of the theory of planned behaviour: A meta-analytic review. British Journal of Social Psychology. 2001;40(4):471–499. PMID: 11795063

Buhrmester, M., Kwang, T., & Gosling, S. (2001) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*, 3-5. doi: 10.1177/1745691610393980

Callans, M. (2012, Oct 19). Validity and Reliability in Testing - What do They Mean? Retrieved October 10, 2017, from https://www.wonderlic.com/blog/reliability-validity-testing-mean/

Cohen, J. (1992). A Power Primer. *Quantitative Methods in Psychology,112*(1), 155-159. doi:0033-2909/92/S3-00

Comrey, A. L., & Lee, H. B. (1992). *A first Course in Factor Analysis*. Hillsdale, NJ: Erlbaum.

Cummings, E. (2009). Trends in mental health googling. *Psychiatric Bulletin, 33*(11), 437-437. doi:10.1192/pb.33.11.437

Flanagin, A. J., & Metzger, M. J. (2011). From Encyclopædia Britannica To Wikipedia. *Information, Communication & Society,14*(3), 355-374. doi:10.1080/1369118x.2010.542823

Godin G, Kok G. The Theory of Planned Behavior: A Review of Its Applications to Health-Related Behaviors. American Journal of Health Promotion. 1996;11:87–98. Doi: 10.4278/0890-1171-11.2.87

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression.* New York: Wiley.

Jenkins, M. M., & Youngstrom, E. A. (2016). A randomized controlled trial of cognitive debiasing improves assessment and treatment selection for pediatric bipolar disorder. *Journal of Consulting and Clinical Psychology,84*(4), 323-333. doi:10.1037/ccp0000070

Jolliffe, I. T. (2002). *Principal component analysis* (151-152). New York: Springer.

Larner, A. J. (2014). Effect Size (Cohen's d) of Cognitive Screening Instruments Examined in Pragmatic Diagnostic Accuracy Studies. *Dementia and Geriatric Cognitive Disorders Extra,4*(2), 236-241. doi:10.1159/000363735

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mchugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *American Psychologist,65*(2), 73-84. doi:10.1037/a0018121

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology,58*(13), 2078-2091. doi:10.1002/asi.20672

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics,59*, 210-220. doi:10.1016/j.pragma.2013.07.012

Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., . . . Jorm, A. F. (2011). Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. *Psychological Medicine,42*(08), 1753-1762. doi:10.1017/s003329171100287x

Richey, M. H., Mcclelland, L., & Shimkunas, A. M. (1967). Relative influence of positive and negative information in impression formation and persistence. *Journal of Personality and Social Psychology,6*(3), 322-327. doi:10.1037/h0024734

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika,41*(3), 321-327. doi:10.1007/bf02293557

Whiting, P., Rutjes A., Reitsma, J., Bossuyt, P., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews**.** *BMC Biomedical Research Methodology, 3*. doi: https://doi.org/10.1186/1471-2288-3-25

Table 1

Demographic differences between quality and not quality with and without the wQUADAS

Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*= 454**	***n*= 477**			
	***n*%**	***n*%**			
		Ratings using the wQUADAS			
Education Level (Bachelor's)	34.9%	34.2%	$X^2(7)=6.19$	0.52	*phi* = 0.07
Gender (Female)	51.9%	58.6%	$X^2(1)=4.58$	0.03*	*phi* = 0.07
Race (White)	70.1%	70.8%	$X^2(5)=3.68$	0.60	*phi* = 0.05
		Ratings without using the wQUADAS			
Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*= 836**	***n*= 174**			
	***n*%**	***n*%**			
Education Level (Bachelor's)	31.6%	37.9%	$X^2(7) = 6.70$	0.46	*phi* = 0.08
Gender (Female)	58.7%	59.1%	$X^2(1) = 0.01$	0.91	*phi* = 0.003
Race (White)	71.6%	58.3%	$X^2(4) = 11.04$	0.03*	*phi* =0.11

**p*<.05, ***p*<.005, ****p*<.0005 two tailed

Table 2

www.prevention.com (Depression) Demographic differences between quality and not quality with and without the wQUADAS

Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*=91**	***n*=75**			
	n%	***n%***			
		Ratings using the wQUADAS			
Education Level	32.3%	35.6%	$X^2(6)=4.58$	0.60	*phi=0.17*
Gender (Female)	43.0%	58.9%	$X^2(1)=4.00$	0.05	*phi* =0.16
Race (White)	75.3%	69.4%	$X^2(5)=7.87$	0.16	*phi=0.22*
		Ratings without using the wQUADAS			
Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*=151**	***n*=31**			
	n%	***n%***			
Education Level	35.6%	45.2%	$X^2(7)=4.06$	0.77	*phi*=0.15
Gender (Female)	55.5%	64.5%	$X^2(1)=0.85$	0.36	*phi*=0.07
Race (White)	70.5%	59.4%	$X^2(4)=1.96$	0.74	*phi*=0.11

**p*<.05, ***p*<.005, ****p*<.0005 two tailed

Table 3

www.wikipedia.com (Major Depression) Demographic differences between quality and not quality with and without the wQUADAS

Variable	Quality	Not Quality	Test statistic	*p*	Effect size
	***n*=105**	***n*=74**			
	***n*%**	***n*%**			
		Ratings using the wQUADAS			
Education Level	36.2%	37.5%	$X^2(7)=5.81$	0.56	*phi=0.18*
Gender (Female)	51.9%	60.6%	$X^2(1)=1.28$	0.26	*phi* =0.09
Race (White)	66.7%	63.4%	$X^2(4)=2.85$	0.58	*phi=0.13*
		Ratings without using the wQUADAS			
Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*=181**	***n*=15**			
	***n*%**	***n*%**			
Education Level	39.5%	43.8%	$X^2(6)=0.81$	0.99	*phi*=0.07
Gender (Female)	59.9%	68.8%	$X^2(1)=0.48$	0.49	*phi*=0.05
Race (White)	78.0%	56.3%	$X^2(4)=4.75$	0.31	*phi*=0.16

*p<.05, **p<.005, ***p<.0005 two tailed

Table 4

www.wikipedia.com (Anxiety) Demographic differences between quality and not quality with and without the wQUADAS

Variable	**Quality** ***n*=52** ***n%***	**Not Quality** ***n*=105** ***n%***	**Test statistic**	***p***	**Effect size**
		Ratings using the wQUADAS			
Education Level	34.6%	32.0%	$X^2(6)=2.84$	0.83	*phi=0.14*
Gender (Female)	48.0%	56.9%	$X^2(1)=1.06$	0.30	*phi* =0.08
Race (White)	65.4%	69.6%	$X^2(4)=4.33$	0.36	*phi=0.17*
		Ratings without using the wQUADAS			
Variable	**Quality** ***n*=151** ***n%***	**Not Quality** ***n*=19** ***n%***	**Test statistic**	***p***	**Effect size**
Education Level	14.2%	20.0%	$X^2(6)=4.76$	0.58	*phi*=0.17
Gender (Female)	59.2%	75.0%	$X^2(1)=1.85$	0.17	*phi*=0.11
Race (White)	78.4%	70.0%	$X^2(4)=6.74$	0.15	*phi*=0.20

*p<.05, **p<.005, ***p<.0005 two tailed

Table 5

www.health.com (Anxiety) Demographic differences between quality and not quality with and without the wQUADAS

Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*=67**	***n*=149**			
	n%	***n%***			
		Ratings using the wQUADAS			
Education Level	37.3%	42.0%	$X^2(6)=4.41$	0.62	*phi=0.15*
Gender (Female)	64.2%	54.8%	$X^2(1)=1.61$	0.20	*phi* =-0.09
Race (White)	68.7%	75.4%	$X^2(4)=2.34$	0.67	*phi=0.11*
		Ratings without using the wQUADAS			
Variable	**Quality**	**Not Quality**	**Test statistic**	***p***	**Effect size**
	***n*=139**	***n*=23**			
	n%	***n%***			
Education Level	41.3%	17.4%	$X^2(6)=16.96$	0.01*	*phi*=0.33
Gender (Female)	62.0%	52.2%	$X^2(1)=0.80$	0.37	*phi*=-0.07
Race (White)	67.4%	69.6%	$X^2(4)=3.97$	0.41	*phi*=0.16

*p<.05, **p<.005, ***p<.0005 two tailed

Table 6

www.webmd.com (suicide) Demographic differences between quality and not quality with and without the wQUADAS

Variable	**Quality** ***n*=62** ***n*%**	**Not Quality** ***n*=107** ***n*%**	**Test statistic**	***p***	**Effect size**
			Ratings using the wQUADAS		
Education Level	33.9%	28.4%	$X^2(6)=6.94$	0.33	*phi=0.21*
Gender (Female)	54.1%	59.8%	$X^2(1)=0.51$	0.48	*phi* =0.06
Race (White)	72.1%	73.5%	$X^2(4)=2.86$	0.58	*phi=0.13*
			Ratings without using the wQUADAS		
Variable	**Quality** ***n*=136** ***n*%**	**Not Quality** ***n*=21** ***n*%**	**Test statistic**	***p***	**Effect size**
Education Level	34.6%	22.2%	$X^2(6)=6.59$	0.36	*phi*=0.21
Gender (Female)	55.5%	50.0%	$X^2(1)=0.26$	0.61	*phi*=-0.04
Race (White)	75.4%	70.4%	$X^2(4)=1.47$	0.83	*phi*=0.10

**p*<.05, ***p*<.005, ****p*<.0005 two tailed

Table 7

www.wikipedia.com (suicide) Demographic differences between quality and not quality with and without the wQUADAS

Variable	**Quality** ***n*=82** ***n*%**	**Not Quality** ***n*=119** ***n*%**	**Test statistic**	***p***	**Effect size**
		Ratings using the wQUADAS			
Education Level (Bachelor's)	35.4%	28.9%	$X^2(7)=16.94$	0.02*	*phi=0.29*
Gender (Female)	51.9%	62.3%	$X^2(1)=2.11$	0.15	*phi*=0.10
Race (White)	69.5%	64.0%	$X^2(4)=1.15$	0.89	*phi=0.08*
		Ratings without using the wQUADAS			
Variable	**Quality** ***n*=78** ***n*%**	**Not Quality** ***n*=65** ***n*%**	**Test statistic**	***p***	**Effect size**
Education Level (Bachelor's)	32.8%	29.2%	$X^2(6)=1.38$	0.97	*phi*=0.10
Gender (Female)	56.0%	66.7%	$X^2(1)=0.92$	0.34	*phi*=0.08
Race (White)	15.9%	28.6%	$X^2(3)=2.22$	0.53	*phi*=0.21

**p*<.05, ***p*<.005, ****p*<.0005 two tailed

Table 8

Linear regression model for wQUADAS score and Logistic regression model for quality rating

Variable	***B***	***SE***	***Beta***	***t***	***p***
Linear regression					
Model 1: Race					
White	0.42	9.44	0.01	0.04	0.97
Black	0.26	9.54	0.01	0.03	0.98
Asian	4.05	10.04	0.04	0.40	0.69
Other	-0.45	9.50	-0.01	-0.05	0.96
Model 2: Gender (female)	-2.0	1.03	-0.06	-1.92	0.06
Model 3: Education	-0.10	0.37	-0.01	-0.26	0.80
Variable	***B***	***SE***	***Wald***	***Exp(B)***	***p***
Logistic regression					
Model 1: Race					
White	0.42	1.23	0.12	1.53	0.73
Black	0.34	1.24	0.07	1.40	0.79
Asian	0.13	1.30	0.01	1.14	0.92
Other	0.44	1.23	0.13	1.55	0.72
Model 2: Gender (female)	-0.27	0.13	4.63	0.76	0.03*
Model 3: Education	-0.07	0.05	2.08	0.94	0.15

*p<.05, **p<.005, ***p<.0005 two tailed