A Standardized Rubric for Evaluating Webquest Design: Reliability Analysis of ZUNAL Webquest Design Rubric

Zafer Unal
USF St. Petersburg, St. Petersburg, FL, USA

Yasar Bodur
Georgia Southern Univ., Statesboro, GA, USA

Aslihan Unal
Uşak Univ., Uşak, Turkey

zafer@zunal.com   yasar@zunal.com   asli@zunal.com

Executive Summary

Current literature provides many examples of rubrics that are used to evaluate the quality of webquest designs. However, reliability of these rubrics has not yet been researched. This is the first study to fully characterize and assess the reliability of a webquest evaluation rubric. The ZUNAL rubric was created to utilize the strengths of the currently available rubrics and improved based on the comments provided in the literature and feedback received from the educators.

The ZUNAL webquest design rubric was developed in three stages. First, a large set of rubric items was generated based on the operational definitions and existing literature on currently available webquest rubrics (version 1). This step included item selections from the three most widely used rubrics created by Bellofatto, Bohl, Casey, Krill & Dodge (2001), March (2004), and eMints (2006). Second, students (n=15) enrolled in a graduate course titled "Technology and Data" were asked to assess the clarity of each item of the rubric on a four-point scale ranging from (1) "not at all" to (4) "very well/very clear." This scale was used only during the construction of the ZUNAL rubric; therefore, it was not a part of the analyses presented in this study. The students were also asked to supply written feedback for items that were either unclear or unrelated to the constructs. Items were revised based on the feedback (version 2,). Finally, K-12 classroom teachers (n=23) that are involved with webquest creation and implementation in classrooms were invited for a survey that asked them to rate rubric elements for their value and clarity. Items were revised based on the feedback.

At the conclusion of this three-step process, the webquest design rubric was composed of nine main indicators with 23 items underlying the proposed webquest rubric constructs: title (4 items), introduction (1 item), task (2 items), process (3 items), resources (3 items), evaluation (2 items), conclusion (2 items), teacher page (2 items) and overall design (4 items). A three-point response scale including "unacceptable", "acceptable", and "target" was utilized.

After the rubric was created, twenty-three participants were given a week to evaluate three pre-selected webquests with varying quality using the latest version of the rubric. A month later, the evaluators were asked to re-evaluate the same webquests.

In order to investigate the internal consistency and intrarater (test retest) reliability of the ZUNAL webquest design rubric, a series of statistical procedures were employed. The statistical analyses conducted on the ZUNAL webquest rubric pointed to its acceptable reliability. It is reasonable to expect that the consistency we observed in the rubric scores was due to the comprehensiveness of the rubric and clarity of the rubric items and descriptors. Because there are no existing studies focusing on reliability of webquests design rubrics, researchers were unable to make comparisons to discuss the merits of the ZUNAL rubric in relation to others at this point.

Keywords: webquest, webquest rubric, rubric reliability analysis, internal consistency, test-retest reliability, interrater reliability

Introduction

The webquest concept was developed in 1995 by Bernie Dodge and Tom March with the purpose of leading the teachers to build educational activities taking advantage of the existing resources on the internet. Rather than simply pointing students to websites which may encourage them to 'copy and paste', well-structured webquests direct students to internet resources that require the use of critical thinking skills and a deeper understanding of the subject being explored (March, 2003).

A webquest is constructed and presented in six parts called building blocks: Introduction, the indication of at least one task, the process through which students will accomplish the task, the Resources where students find the information needed to accomplish the task, the evaluation used to assess the product resulting from students' work, and a conclusion. An additional section is called teacher page in which webquest designers provide detailed information regarding webquests such as curriculum standards, credits, and worksheets (Dodge, 1999). For example, Figure 1 features a

Figure 1. A sample webquest from zunal: http://zunal.com/webquest.php?w=4603.

webquest activity in which students are asked to plan a one-week vacation based on a fixed budget with certain requirements, number of things to do, number of places to visit, etc. At the end of the webquest activity, students are asked to present their vacation plan to the class.

According to Dodge (1995, 1997) the webquests can be short term (between one and three classes) or long term (from one week to a month). A short term webquest is focused on the acquisition and integration of a certain amount of knowledge by the student. A long term webquest allows a deeper analysis of concepts as well as broadens and reinforces the acquired knowledge.

The way a webquest activity is designed discourages students from simply surfing the Web in an unstructured manner. In a webquest, students are given a set of internet sites to visit to collect information about a specific topic related to the class curriculum. The teacher gives the class the process to follow and the desired result. Once students have collected the necessary information, they evaluate their findings and report to the teacher. As a conclusion, the class reviews the findings and report to the teacher. One thing to keep in mind that Dodge (1995) cautions educators is not to confuse webquest with "scavenger or treasure hunt" activities. In a scavenger or treasure hunt the questions are predetermined and the answers are static. In a webquest the students are given a task in which they will be required to use the information they have. If the task is a scavenger hunt it is not a webquest. A webquest is about what they do with the information collected from the resources.

Teachers report that the experience of designing and implementing webquests helps them "discover new resources, sharpen technology skills, and gain new teaching ideas by collaborating with colleagues" (Peterson & Koeck, 2001, p. 10). Since webquests challenge student intellectual and academic ability rather than their simple web searching skills, they are said to be capable of increasing student motivation and performance (March, 2004), developing students' collaborative and critical thinking skills (Perkins & McKnight, 2005; Tran, 2010), and enhancing students' ability to apply what they have learned to new learning (Pohan & Mathison, 1998; Tran, 2010).

Many research studies have been conducted to determine the effects of webquests on teaching and learning in many different disciplines and grade levels. It has been indicated that webquest activities create positive attitudes and perceptions among students (Gorrow, Bing, & Royer, 2004; Tsai, 2006; Unal & Leung, 2010), increase the learners' motivation (Abbit & Ophus, 2008; Carneiro & Carvalho, 2011; Tsai, 2006), foster collaboration (Barroso & Coutinho, 2010; Bartoshesky & Kortecamp, 2003), enhance problem-solving skills, higher order thinking, and connection to authentic contexts (Abu-Elwan, 2007; Allen & Street, 2007; Lim & Hernandez, 2007), improve students' reading abilities (Chou, 2011), and assist in bridging the theory to practice gap (Laborda, 2009; Lim & Hernandez, 2007). In addition, teachers reported positively to the value of the webquests in their daily teaching (Oliver, 2010; Yang, Tzuo & Komara, 2011).

Considering its increasing use, quality of webquests is an important matter. While webquests show great promise in enhancing student learning and motivation, the results of using webquests as teaching and learning tools may depend on how well webquests are designed in the first place. There are thousands of webquests on the internet but the quality of these webquests varies (Dodge, 2001; March, 2003). As a matter of fact, some of them may not be considered as real webquests (March, 2003). March claims that a good webquest must be able to "prompt the intangible 'aha' experiences that lie at the heart of authentic learning" (March, 2003, p. 42). Both Dodge and March indicate that a careful evaluation is needed before adapting a webquest to be used in classroom with students (Dodge, 2001; March, 2003). Therefore, careful and comprehensive evaluation of webquest design is an essential step in the decisions of using webquests. Rubrics are "one way to evaluate whether a webquest is well designed. A rubric is a rule or a set of rules or directions for doing an action as a ritualistic part of a situation (Skovira, 2009). Using rubrics with scoring criteria can provide meaningful assessment information (Buzzetto-More &

Alade, 2006; Petkov & Petkova, 2006). Providing these rubrics in digital format offered on the web and connected to a database, the rubrics provide educators with data that can be aggregated (Buzzetto-More, 2006).

There are many existing webquest rubrics that are used to judge whether a webquest is well designed however only three rubrics are widely used.

Rubric 1. Rubric for Evaluating Webquests by Dodge (1997)

http://webquest.sdsu.edu/webquestrubric.html

Dodge (1997) created a rubric for evaluating webquests, which was advanced in Bellofatto et al. (2001). The rubric is designed to evaluate the overall aesthetics, as well as the basic elements of a webquest. Every category is evaluated according to 3 levels; Beginning, Developing, and Accomplished. Every cell is worth a number of points. The teacher can take advantage of all the opportunities afforded on the rubric, score every dimension of the webquest, and come up with a score out of a total of 50 points that objectifies the usefulness of a webquest (Bellofatto et al., 2001). While this is the most commonly used webquest design evaluation rubric, there have been discussions and suggestions regarding the certain elements of it. For example, Maddux and Cummings (2007) discussed the lack of focus on the "learner" and recommended the addition of "learner characteristics" to the "Rubric for Evaluating webquests" (Bellofatto et al., 2001). They asserted that "…the rubric did not contain any category that would direct a webquest developer to consider any characteristics of learners, such as ages or cognitive abilities. Instead, the rubric focused entirely on the characteristics of the webquest which does nothing to ensure a match between webquest's cognitive demands and learner characteristics, cognitive or otherwise" (p. 120). Finally they suggested that care should be taken to ensure that teachers who develop and use webquests are mindful of students' individual differences including, but not limited to, age, grade, and cognitive developmental level. To remind teachers of the importance of these considerations, Dodge's (1997) second item in his list of webquests's critical attributes should be modified from "a task that is doable and interesting" to "a task that is doable, interesting, and appropriate to the developmental level and other individual differences of students with whom the webquest will be used" (p. 124).

Rubric 2. Webquest Assessment Matrix by March (2004)

http://bestwebquests.com/bwq/matrix.asp

March (2004) created a rubric for evaluating webquest design called Webquest Assessment Matrix. The rubric has eight criteria (Engaging Opening / Writing, the Question/task, background for everyone, Roles/Expertise, Use of the Web, Transformative Thinking, Real World Feedback, and Conclusion). One unique aspect of this evaluation rubric is that it does not have specific criteria for web elements such as graphics and web publishing. March (2004) suggests that one person's cute animated graphic is another's flashing annoyance. There are 8 categories and every category is evaluated according to 3 aspects; Low (1 point), Medium (2 points), and High (3 points). The maximum score is 24 points. March (2004) also offers a subscription based webquest personalized evaluation/feedback on his official website with the use of this assessment matrix.

Rubric 3. eMINTS Rubric (2006)

http://www.emints.org/inside-emints/webquests

eMINTS (enhancing Missouri's Instructional Networked Teaching Strategies) National Center also created a rubric based on Dodge's work (eMINTS, 2006). Webquest creators are asked to use this rubric to evaluate their webquest design before submitting it for eMINTS evaluation. The

eMINTS National Center then evaluates submissions and provides a link to webquests on their website that score 65 points or higher (Total 70 points possible). This provides authenticity for webquest creators.

As in every assessment tool, a webquest assessment rubric should be independent of who does the scoring and the results similar no matter when and where the evaluation is carried out. The more consistent the scores are over different raters and occasions, the more reliable the assessment is thought to be (Moskal & Leydens, 2000). The current literature provides examples of rubrics that are used to evaluate the quality of webquest design. However, in the current literature, the discussion of reliability analyses of webquest design rubrics is non-existent. This study focuses on reliability analysis of a webquest design evaluation rubric.

This article is structured as follows: First a brief overview of rubric reliability and details on different forms of reliability calculations are presented. In the next section, authors describe the specific details on the procedure followed in this study. Next in the results section, the different reliability calculations (internal consistency and intrarater (test-retest)) of the new webquest rubric are provided. The last section outlines the conclusions and future work.

Reliability of Rubrics

There is "nearly universal" agreement that reliability is an important property in educational measurement (Colton, Gao, Harris, Kolen, Martinovich-Barhite, Wang, & Welch, 1997, p. 3). Many assessment methods require raters to judge or quantify some aspect of student behavior (Stemler, 2004), and Johnson, Penny & Gordon (2000) challenge those who design and implement assessments to strive to achieve high levels of reliability. Ideally, an assessment should be independent of who does the scoring and the results should be similar no matter when and where the assessment is carried out, but this is hardly attainable. The more consistent the scores are over different raters and occasions, the more reliable the assessment is considered (Moskal & Leydens, 2000).

Two forms of reliability are considered significant. The first form is interrater reliability, which refers to the consistency of scores assigned by multiple raters, while the second is intrarater reliability, which refers to the consistency of scores assigned by one rater at different points of time (Moskal, 2000).

Interrater reliability refers to "the level of agreement between a particular set of judges on a particular instrument at a particular time" and "provide[s] a statistical estimate of the extent to which two or more judges are applying their ratings in a manner that is predictable and reliable" (Stemler, 2004, p. 3). Raters, or judges, are used when student products or performances cannot be scored objectively as right or wrong but require a rating of degree (Stemler, 2004). This use of raters results in the subjectivity that comes hand in hand with a rater's interpretation of the product or performance (Stemler, 2004). Perhaps the most popular statistic for calculating the degree of consistency between judges is the Pearson correlation coefficient (Stemler, 2004). One beneficial feature of the Pearson correlation coefficient is that the scores on the rating scale can be continuous in nature (e.g., they can take on partial values such as 1.5). Like the percent-agreement statistic, the Pearson correlation coefficients can be calculated only for one pair of judges at a time and for one item at a time. Values greater than .70 are typically acceptable for consistency estimates of interrater reliability (Barrett, 2001; Glass & Hopkins, 1996; Stemler, 2004). In situations where multiple judges are used, another approach to computing a consistency estimate of interrater reliability would be to compute Cronbach's alpha coefficient (Crocker & Algina, 1986). Cronbach's alpha coefficient is used as measure of consistency when evaluating multiple raters on ordered category scales (Bresciani, Zelna, & Anderson, 2009). If the Cronbach's alpha estimate is low, then this implies that the majority of the variance in the total composite score is

really due to the error variance, and not the true score variance (Crocker & Algina, 1986). The interpretation of Cronbach's alpha is that it is the expected correlation between pairs of student scores if we were to choose two random samples of judges and compute two different scores for each student each based on the judges. Though some authors discourage the assignment of strength of reliability scale to this statistic as it is dependent on the number of judges (Cortina, 1993), 0.7 is generally considered a satisfactory value of alpha (Nunnally, 1978).

Intrarater (test retest) reliability refers to the consistency of scores assigned by one rater at different points of time (Speth, Namuth, & Lee, 2007; Moskal & Leydens, 2000). Unlike measures of internal consistency that provide the extent to which all of the questions that make up a scale measure the same construct, measures of temporal stability tell you whether or not the instrument is consistent over time and/or over multiple administrations. The test is performed twice. In the case of a rubric, this would mean evaluating subjects using the same rubrics by the same group of evaluators on two different occasions. If the correlation between separate administrations of the evaluation is high, then it is considered to have good test-retest reliability. The test-retest reliability coefficient is simply a Pearson correlation coefficient for the relationship between the total scores for the two administrations. Additionally, intraclass correlation coefficient (ICC) is used when consistency between ratings from the same raters are evaluated.

The following section describes the process through which the webquest evaluation rubric was created. After that, results of reliability analyses on the rubric are stated.

Procedures

The rubric was developed in three stages. First, a large set of rubric items was generated based on the operational definitions and existing literature on currently available webquest rubrics (version 1). This step included item selections from the three most widely used rubrics created by Bello-fatto et al., (2001), March (2004), and eMints (2006). Second, students (n=15) enrolled in a graduate course titled "Technology and Data" were asked to determine the clarity of each item on a four-point scale ranging from (1) "not at all" to (4) "very well/very clear." They were also asked to supply written feedback for any items that were either unclear or unrelated to the constructs. Items were revised based on the feedback (version 2,). Finally, K-12 classroom teachers (n=23) that are involved with webquest creation and implementation in classrooms were invited for a survey that them asked to rate rubric elements for their value and clarity. Items were revised based on the feedback (final version, Appendix 1). At the conclusion of this three-step process, the webquest evaluation rubric was composed of nine subscales with 23 items underlying the proposed webquest rubric constructs: title (4 items), introduction (1 item), task (2 items), process (3 items), resources (3 items), evaluation (2 items), conclusion (2 items), teacher page (2 items) and overall design (4 items). A 3-point response scale including "unacceptable", "acceptable", and "target" was utilized (see Appendix 1). The authors realized the limitations of having only one item for the "introduction" construct for the reliability analysis. However, the literature indicated that the "introduction" section on a webquest plays the role only of motivating students into the "task" by providing effective background information on the topic. Therefore, the authors did not add any more items in this section order to keep the overall scoring consistent based on the importance of different webquest components and their weight in the scoring.

ZUNAL (www.zunal.com) is a web application available on the internet for educators to create and publish webquests. It currently has more than 138,000 members and receives approximately 300 new users daily. Currently, there are over 54,000 published webquests available for public. The researchers selected to use this web application for the study purposes in order to find select webquests and evaluators for reliability testing.

Three webquests were selected for evaluation by the researchers in order to investigate the reliability of the rubric from the web application, ZUNAL. Fortunately, ZUNAL already has a rating system where users and visitors post their ratings on a five-star rating basis and comments for each webquest. In order to ensure the inclusion of webquests with varying quality, researchers randomly selected a low quality (1 and 2 stars), a medium quality (3 and 4 stars), and a high quality (5 stars) webquest in the study (Webquest 1: low quality, Webquest 2: medium quality, and Webquest 3: high quality). The selected webquests were all in different topics: social studies, English/language arts, and science respectively. All three webquests were located at ZUNAL (www.zunal.com).

After the integration and testing of the rubrics at the ZUNAL webquest builder website and the selection of the three webquests with varying quality, an online invitation survey was designed and sent to ZUNAL users. The survey provided the description of the study, consent form, and questions for determining the eligibility for participation in the study. Only users that already had a webquest published at ZUNAL with a higher than 3-star review were selected and invited. Of the 118 selected users, 23 agreed to participate in the study.

Seven males (30%) and 16 females (70%) participated in the study, and their ages varied from 26 to 53 (M = 39.34, SD=7.95). 11 participants were K-12 teachers, 8 participants were college faculty from both public and private universities, and finally 4 participants were College of Education graduate students from various colleges. As stated above, all of the participants had published a webquest with 4- or 5-star rating on ZUNAL webquest builder site.

The 23 participants were given a week to evaluate the three webquests using the latest version of the rubric. A month later, the evaluators were asked to re-evaluate the same webquests again. In order to encourage their participation, researchers offered to upgrade their ZUNAL account for unlimited webquest creation. Researchers selected a one month (30 days) time span between the evaluations because choosing less than a month (such as a week) would have led to high correlation scores. However, researchers believed that extending to longer than one month would have led to a lower return rate since these were voluntary evaluations with account promotion offer. The rubric scores on each webquest were entered into SPSS 17.0 for reliability analysis.

Results

In order to investigate the internal consistency, intrarater (test retest) reliability, and interrater reliability of the ZUNAL webquest evaluation rubric, a series of statistical procedures were employed. Before the analyses, all ratings had been quantified using the following point values: unacceptable 1 point, acceptable 2 points, and target 3 points.

First we averaged all raters' ratings on each webquest and assigned an overall mean score to first and second ratings of each webquest. Second, we correlated the mean scores for the first and second ratings of all three webquest. Table 1 provides descriptive information about the first and second ratings of each webquest.

Table 1: Descriptive information on first and second ratings			
Rating	N	Mean	Standard Deviation
WQ1 First rating	23	1.93	.054
WQ1 Second rating	23	1.96	.055
WQ2 First rating	23	2.48	.054
WQ2 Second rating	23	2.51	.074
WQ3 First rating	23	2.74	.083
WQ3 Second rating	23	2.76	.079

Rating Scale Unacceptable (1)   Acceptable (2)   Target (3)

As the table indicates, the first and second rating scores calculated based on all raters' ratings were very similar with small standard deviations in comparison to the mean scores.

Internal Consistency

First, Cronbach's alpha was calculated for the first and second rating of all three webquests to assess the internal consistency of the webquest evaluation instrument. Cronbach's alpha reliability coefficient normally ranges between 0 and 1. However, there is actually no lower limit to the coefficient. The closer Cronbach's alpha coefficient is to1.0 the greater the internal consistency of the items in the scale. In this study, Cronbach's alpha on the first and second ratings of the webquests using the aforementioned rubric was .956 and .953 respectively indicating a high level of internal consistency in the evaluations.

Intrarater (Test Retest) Reliability

A Pearson correlation was calculated to examine test-retest consistency between the first and second ratings of each webquest as presented in Table 2. As the Table 2 indicates, averaged first and second ratings on each webquest were positively correlated and significant at the 1% level.

Table 2: Pearson correlations on averaged first and second ratings			
Rating	**Webquest 1 Second Rating**	**Webquest 2 Second Rating**	**Webquest 3 Second Rating**
Webquest 1First Rating	.851**		
Webquest 2First Rating		.606**	
Webquest 3First Rating			.860**

** Correlations significant at the 1% level.

Moreover, a Pearson correlation was conducted as a test-retest reliability measure to examine the consistency between the first and second ratings of each evaluator on each webquest. These results are presented in Table 3.

Table 3: Pearson correlations on individual rater's first and second ratings			
Rater / Webquest	**Webquest 1**	**Webquest 2**	**Webquest 3**
1	.875**	.878**	1.00**
2	.880**	.820**	.872**
3	.851**	.829**	.913**
4	.937**	1.00**	.872**
5	.801**	1.00**	.900**
6	.821**	.807**	1.00**
7	.904**	.940**	.907**
8	.938**	.842**	.798**
9	1.00**	1.00**	.907**
10	1.00**	.939**	1.00**
11	1.00**	.940**	.889**
12	.937**	.832**	.692**
13	1.00**	.942**	.917**
14	1.00**	1.00**	.799**
15	.881**	.939**	.900**
16	.929**	.845**	1.00**
17	.938**	.885**	1.00**
18	1.00**	.940**	.889**

19	.929**	.940**	1.00**
20	.938**	.920**	.889**
21	.937**	.939**	.799**
22	.937**	.875**	.845**
23	1.00**	.729**	.697**

**. Correlation is significant at the 1% level (2-tailed)

As the results indicate, the evaluators' one month apart two evaluations of the same webquest yielded very high and significant correlations ranging from .697 to 1.00. As additional measure of test-rests reliability, intraclass correlation coefficients were calculated on each evaluation of each webquest.

The last step of intrarater reliability analysis was conducted on each of the 23 indicators of the rubric using a Pearson correlation on each of the webquests. The analysis indicated high decision consistency. The raters' first and second ratings of each rubric item were significantly correlated at the 1% level for each webquest with the following ranges of correlation coefficients: webquests 1: .73 to 1.00; webquest 2: .69 to 1.00; and webquest 3: .80 to 1.00.

Interrater Reliability

In order to assess the interrater reliability of the ZUNAL webquest evaluation rubric, interclass correlation coefficients for single measures were calculated: webquest 1-evaluation 1: .825; webquest 1-evaluation 2: .797; webquest 2-evaluation 1: .793; webquest 2-evaluation 2: .760; webquest 3-evaluation 1: .673; and webquest 3-evaluation 2: .766. Each of these correlations was significant at the 1% level.

In addition, the authors examined the agreement on the scores assigned to each rubric item by the raters. Table 4 presents the frequencies of scores on each rubric item as percentages. As the table indicates, raters agreed on their assessment of individual items with small variation. On most of the items, more than 90 percent of the raters assigned the same score.

In addition to the frequency analysis, we examined whether the raters' ratings correlated with one another using Pearson correlation. For this specific analysis, we used the ratings on the first scoring of the first and second webquests. This analysis yielded a large number of correlations as we analyzed the relationships among all 23 raters' rating. The analysis indicated the on the first rating of the first webquest, all 23 raters' scores significantly correlated with each other at the 1% level. The strength of the correlations ranged from .602 to 1.00. On the first administration of the second webquest, all the ratings were significantly correlated at the 1% level except for one correlation which was significant at the 5% level. The strength of the correlations ranged from .507 to 1.00.

Both the frequencies and the correlation results indicate that the ratings showed inter-rater reliability. Although the correlations showed a range from moderately strong to very strong, most of the correlations were .80 or higher indicating a strong inter-rater agreement.

Table 4: Frequency distribution of scores on webquests as percentages

Item/score	First webquest						Second webquest						Third webquest					
	First rating			Second rating			First rating			Second rating			First rating			Second rating		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	0	21.7	78.3	0	21.7	78.3	0	0	100	0	0	100	0	0	100	0	0	100
2	0	91.3	8.7	0	91.3	8.7	0	0	100	0	0	100	0	0	100	0	0	100
3	0	91.3	8.7	0	91.3	8.7	0	0	100	0	0	100	0	0	100	0	0	100
4	0	87	13	0	87	13	0	0	100	0	0	100	0	0	100	0	0	100
5	87	13	0	78.3	21.7	0	0	95.7	4.3	0	91.3	8.7	0	0	100	0	0	100
6	0	100	0	0	100	0	0	47.8	52.2	0	47.8	52.2	0	0	100	0	0	100
7	100	0	0	100	0	0	0	87	13	0	87	13	0	34.8	65.2	0	30.4	69.6
8	4.3	95.7	0	0	100	0	0	95.7	4.3	0	100	0	0	43.5	56.5	0	43.5	56.5
9	95.7	4.3	0	91.3	8.7	0	0	78.3	21.7	0	78.3	21.7	0	34.8	65.2	0	34.8	65.2
10	13	87	0	13	87	0	0	30.4	69.6	0	26.1	73.9	0	0	100	0	0	100
11	26.1	73.9	0	26.1	65.2	8.7	0	26.1	73.9	0	30.4	69.6	0	0	100	0	0	100
12	8.7	91.3	0	0	100	0	0	91.3	8.7	0	91.3	8.7	0	69.6	30.4	0	65.2	34.8
13	8.7	91.3	0	0	100	0	0	100	0	0	87	13	0	52.2	47.8	0	47.8	52.2
14	4.3	95.7	0	0	100	0	0	100	0	0	87	13	0	34.8	65.2	0	34.8	65.2
15	4.3	95.7	0	0	100	0	0	95.7	4.3	0	91.3	8.7	0	95.7	4.3	0	69.6	30.4
16	4.3	95.7	0	0	100	0	0	95.7	4.3	0	91.3	8.7	0	52.2	47.8	0	52.2	47.8
17	4.3	95.7	0	4.3	95.7	0	0	0	100	0	4.3	95.7	0	0	100	0	0	100
18	100	0	0	73.9	26.1	0	78.3	21.7	0	47.8	39.1	13	0	100	0	0	82.6	17.4
19	100	0	0	0	100	0	0	26.1	73.9	0	26.1	73.9	0	4.3	95.7	0	4.3	95.7
20	0	4.3	95.7	0	4.3	95.7	0	0	100	0	0	100	0	0	100	0	4.3	95.7
21	0	100	0	0	100	0	0	100	0	0	91.3	8.7	0	82.6	17.4	0	82.6	17.4
22	0	100	0	0	100	0	0	0	100	0	95.7	0	0	0	100	0	0	100
23	0	0	100	0	0	100	0	0	100	0	0	100	0	4.3	95.7	0	0	100
24	0	100	0	0	100	0	0	0	100	0	0	100	0	0	100	0	0	100

Summary and Conclusions

Summary

The researchers in this study undertook development of the ZUNAL webquest design rubric and investigated its reliability using multiple measures. To our knowledge, this is the first study to fully assess the reliability of a webquest design evaluation rubric. The rubric was created to use the strengths of the currently available rubrics and improved based on the comments provided in the literature and feedback received from the educators. The outcome of the process was a very comprehensive rubric that took into account the technical, pedagogical, and aesthetic aspects of webquests as well as characteristics of learners. Because webquests are used in schools and class-rooms for educational purposes, we believe that a rubric should take into account not only the subject matter and its presentation, but also the characteristics of the learners who will work with the webquest.

The statistical analyses conducted on the ZUNAL webquest design rubric pointed to its accept-able reliability. Internal consistency using the Cronbach's alpha on the first and second ratings of the webquests was .956 and .953 respectively showing a high level of consistency. When the data were analyzed of the intrarater (test-retest) reliability, a Pearson correlation calculation between the first and second ratings of each webquest were positively correlated and significant at the 1% level. In addition, a Pearson correlation was also used to measure the consistency between the first and second ratings of each evaluator on each webquest. The results showed very high and significant correlations ranging from .697 to 1.00. Moreover, intraclass correlation coefficients were also calculated on each evaluation of each webquest and each of these correlations was also found significant at the 1% level. Finally, additional intrarater reliability analysis conducted on each of the 23 rubric indicators using a Pearson correlation yielded a high decision consistency (webquests 1: .73 to 1.00; webquest 2: .69 to 1.00; and webquest 3: .80 to 1.00.). For the inter-rater reliability, interclass correlation coefficients for single measures ranged between .673 and .825 and were found significant at the 1% level. In addition, agreement on the scores assigned to each rubric item by the raters was calculated. Raters agreed on their assessment of individual items with small variation. On most of the items, more than 90 percent of the raters assigned the same score.

It is reasonable to expect that the consistency we observed in the rubric scores was due to the comprehensiveness of the rubric and clarity of the rubric items and descriptors. However, it is also possible that the limited range of the scale used may have inflated the correlations. Because there are no existing studies focusing on reliability of webquests design rubrics, researchers were unable to make comparisons discussions at this point.

It is important to note that two of the elements (title and keywords) in the given rubric seem to be common elements of a webquest design process that might seem to be easily achievable, there-fore leading to high reliability scores. Researchers and other educators participated in the study considered these elements very carefully, constructing sophisticated criteria for each because they are as crucial as other elements of the rubric. To illustrate, an educator who created a webquest on the lifecycle of a butterfly can name the webquest "lifecycle of a butterfly" or "lifecycles" or "butterflies" or perhaps "A life science webquest." Only the first title points to the content of the webquests; therefore, only the first title would be considered acceptable according to the ZUNAL rubric. This is especially important in the context of educators searching for and using webquests available on the internet. Without relevant titles, educators can spend much longer time than nec-essary to find what they are looking for.

Conclusion

The researchers conclude that the final version of the ZUNAL webquest rubric holds promise as an assessment tool for evaluating webquests. As a matter of fact, ZUNAL webquest maker has officially adopted the rubric on its web application and it is currently being used officially by thousands of users.

As there are limitations with all research, this study is no exception. This study is limited to 23 evaluators where each evaluated three webquests at various quality levels. Increasing the number of participants and number of webquests can enhance the generalization of the results. Secondly, webquests selected for evaluation were selected from a single source (www.zunal.com). Choosing webquests that are not template based from other sources can also help increase the generalization. Thirdly, this rubric only evaluates the design of the webquest. It does not measure the learning process from the webquests and content learning. And lastly, our reliability analyses evaluated the ZUNAL webquest design rubric's internal consistency, interrater reliability, and intrarater reliability.

The authors believe that this study provides a strong webquest design rubric with initial reliability data analysis. Further studies are warranted to replicate the current study, investigate webquest evaluations using different rubrics, and assess the relationship between the quality of the webquest design and the extent of student learning.

References

Abbit, J., & Ophus, J. (2008). What we know about the impacts of webquests: *A Review of Research*. *AACE Journal, 16*(4), 441-456.

Abu-Elwan, R. (2007). The use of webquest to enhance the mathematical problem-posing skills of preservice teachers. *International Journal for Technology in Mathematics Education*, *14*(1), 31-39.

Allen, J., & Street, M. (2007). The quest for deeper learning: An investigation into the impact of a knowledge pooling webquest in primary initial teacher training. *British Journal of Educational Technology, 38*(6), 1102-1112.

Barrett, P. (2001). *Assessing the reliability of rating data*. Retrieved July 11, 2011, from http://www.liv.ac.uk/~pbarrett/rater.pdf

Barroso, M., & Coutinho, C. (2010). A webquest for adult learners: A report on a biology course. In J. Sanchez & K. Zhang (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 1566-1569). Chesapeake, VA: AACE.

Bartoshesky, A., & Kortecamp, K. (2003). Webquest: An instructional tool that engages adult learners, promotes higher level thinking and deepens content knowledge. In C. Crawford et al. (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1951-1954). Chesapeake, VA: AACE.

Bellofatto, L., Bohl, N., Casey, M., Krill, M., & Dodge, B. (2001). *A rubric for evaluating webquests*. Retrieved July 11, 2011, from http://webquest.sdsu.edu/webquestrubric.html

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). *Techniques for assessing student learning and development: A handbook for practitioners.* Washington, DC: NASPA.

Buzzetto-More, N., (2006). The e-Learning and business education paradigm: Enhancing education, assessment, and accountability. *Proceedings of the Maryland Business Education Association Conference,* Ocean City, MD.

Buzzetto-More, N., & Alade, A. J., (2006). Best practices in e-assessment. *Journal of Information Technology Education, 5*, 251-269. Retrieved from http://www.jite.org/documents/Vol5/v5p251-269Buzzetto152.pdf

Carneiro, A., & Carvalho, A. A. (2011). A study about WebQuest: The importance of the process in group work. In T. Bastiaens & M. Ebner (Eds.), *Proceedings of World Conference onEducational Multimedia, Hypermedia and Telecommunications 2011* (pp. 3543-3548). Chesapeake, VA: AACE. Retrieved July 11, 2011, from http://www.editlib.org/p/38369

Chou, C. (2011). An inquiry into the effects of incorporating webquest in an EFL college course. *Proceedings of Consumer Electronics, Communications and Networks Conference 2011* (pp.1-4). CECNet: XianNing

Colton, D. A., Gao, X., Harris, D. J., Kolen, M. J., Martinovich-Barhite, D., Wang, T., & Welch, L. (1997). *Reliability issues with performance assessments: A collection of papers*. ACT Research Report Series 97-3.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Dodge, B. (1995). *Some thoughts about webquests*. Retrieved July 11, 2011, from http://edweb.sdsu.edu/courses/edtec596/aboutwebquests.html

Dodge, B. (1997). *A rubric for evaluating webquests*. Retrieved July 11, 2011, from http://webquest.sdsu.edu/webquestrubric.html

Dodge, B. (1999). *Webquest taxonomy: A taxonomy of tasks*. Retrieved July 11, 2011 from http://edweb.sdsu.edu/WebQuest/taskonomy.html

Dodge, B. (2001). FOCUS: Five rules for writing a great webquest. *Learning and Leading with Technology, 28*(8), 6 9, 58.

eMints. (2006). *Rubric/scoring guide for webquests*. Retrieved July 11, 2011, from http://emints.org/webquest/rubric.shtml

Glass, G. V., & Hopkins, K. H. (1996). *Statistical methods in education and psychology*. Boston: Allyn and Bacon.

Gorrow, T., Bing, J., & Royer, R. (2004). *Going in circles: The effects of a webquest on the achievement and attitudes of prospective teacher candidates in education foundations*. Paper presented at the Society for Information Technology and Teacher Education International Conference 2004, Atlanta, GA.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and inter-rater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13,* 121–138.

Laborda, J. G. (2009). Using webquests for oral communication in English as a foreign language for tourism studies. *Educational Technology & Society, 12*(1), 258–270.

Lim, S., & Hernandez, P. (2007). The webquest: An illustration of instructional technology implementation in MFT training. *Contemporary Family Therapy, 29*, 163-175.

Maddux, C. D., & Cummings, R. (2007). Webquests: Are they developmentally appropriate? *The Educational Forum, 71*(2), 117-127.

March, T. (2003). The learning power of webquests. *Educational Leadership, 61*(4), 42-47.

March, T. (2004). *Criteria for assessing best webquests*. Retrieved July 11, 2011, from http://bestwebquests.com/bwq/matrix.asp

Moskal, B. M. (2000). Scoring rubrics: What, when, and how? *Practical Assessment Research and Evaluation, 7*(3). Retrieved July 11, 2011, from http://pareonline.net/getvn.asp?v=7&n=3

Moskal, B., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved July 11, 2011, from http://pareonline.net/getvn.asp?v=7&n=10

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Oliver, D. (2010). *The effect and value of a WebQuest activity on weather in a 5th grade classroom*. Pro-Quest Dissertations and Theses; Thesis (Ed.D.)--Idaho State University, 2010; Publication Number: AAI3405042; ISBN: 9781109700404; Source: Dissertation Abstracts International, Volume: 71-04, Section: A, page: 1274.

Perkins, R., & McKnight, M. L. (2005). Teachers' attitudes toward webquests as a method of teaching. *Computers in the Schools, 22*(1/2), 123-133.

Peterson, C. L., & Koeck, D. C. (2001). When students create their own webquests. *Learning and Leading with Technology, 29*(1), 10–15.

Petkov, D., & Petkova, O., (2006). *Development of scoring rubrics for IS projects as an assessment tool. Proceedings of Informing Science and Information Technology Education Conference* (pp.499 - 510). Manchester, England. Retrieved from http://proceedings.informingscience.org/InSITE2006/IISITPetk214.pdf

Pohan, C., & Mathison, C. (1998). Webquests: The potential of internet based instruction for global education. *Social Studies Review, 37*(2), 91-93.

Skovira, R. J., (2009). Modeling an assessing rubric: Reflections of red ink. *Issues in Informing Science and Information Technology Education, 6*, 437 - 448. Retrieved from http://iisit.org/Vol6/IISITv6p437-448Skovira677.pdf

Speth, C. A. Namuth, D. A., & Lee, D. J. (2007). Using the ASSIST short form for evaluating an information technology application: validity and reliability issues. *Informing Science: The International Journal of an Emerging Transdiscipline, 10*(1), 107-119. Retrieved July 11, 2011, from http://www.inform.nu/Articles/Vol10/ISJv10p107-119Speth104.pdf

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9.

Tran, T. (2010). Using webquest in teaching environmental education in Vietnam. In D. Gibson & B. Dodge (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2010* (pp. 3740-3744). Chesapeake, VA: AACE.

Tsai, S. (2006). *Students' perceptions of English learning through EFL webquest*. Paper presented at The World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006, Orlando, FL.

Unal, Z., & Leung, C. (2010). Identifying preservice and inservice teachers' conceptions of using webquests for classroom instruction via ZUNAL WebQuest Maker. In J. Sanchez & K. Zhang (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2010* (pp. 2748-2757). Chesapeake, VA: AACE. http://www.ncsu.edu/meridian/jul99/webquest/index.htm

Yang, C., Tzuo, P., & Komara, C. (2011). Using webquest as a universal design for learning tool to enhance teaching and learning in teacher preparation programs. *Journal of College Teaching & Learning, 8*(3).

Biographies

Zafer Unal is an Assistant Professor of Childhood Education, serving as assessment coordinator for the College of Education, coordinating the college wide assessment system at University of South Florida, St. Petersburg. Dr. Unal earned his doctorate degree in Elementary Education from Florida State University. His areas of research are using technology as an instructional tool, distance education, and assessment of teacher candidate learning and online portfolios. He teaches undergraduate and graduate courses.

Yasar Bodur is an Associate Professor of Elementary Education, teaching undergraduate and graduate courses for the College of Education at Georgia Southern University. Dr. Bodur earned his doctorate degree in Elementary Education from Florida State University. His areas of research are preservice teacher preparation in multicultural education, field experience in teacher education, and peer coaching.

Asli Unal is an Assistant Professor of Elementary Education, teaching undergraduate and graduate courses for the College of Education at Usak University, Turkey. Dr. Unal earned her doctorate degree in Elementary Education from Florida State University. Her areas of research are parental involvement, educational technology and classroom management.