National Science Teachers' Association

Webwatchers Rubric Analysis

Dr. Maurya Schweizer
August 4, 2003

Introduction

This pilot study seeks to determine the reliability of eight Webwatcher rubrics used for evaluating educational websites. Based on the information obtained from this study, the rubrics will be revised and a second round of data collection will occur to determine if the reliability of the rubrics has increased.

The initial rubrics contained from two to five levels of quality. However, the decision was made to revise the rubrics so that each contains three levels of quality. This was important from a statistical standpoint as interitem correlations and interitem reliability analysis cannot be run on scales that contain both dichotomous and multi-level items.

Overview of Data Collection

A total of 15 middle grade teachers were invited to participate in the study. Two websites aligned with the middle school curriculum were selected for these teachers to evaluate. The participants were asked to rank the website according to the criteria defined in the rubric and to enter comments on why they thought the ranking they chose was the most appropriate. The data collection was done in three phases.

- **Phase One:** Teachers were asked to evaluate the Curve Ball website
- **Interim Phase:** Teachers were asked to re-evaluate the Curve Ball website using only the three rubrics that were modified based on data from Phase One.
- **Phase Two:** Teachers were asked to evaluate the Bucket Buddies website using the revised rubrics.

Phase 1:

Phase One provided quantitative data to assess the reliability of the rubrics, and qualitative data to give insight into potential weaknesses in the rubric and how the rubrics could be modified to improve reliability. Of the 15 teachers invited to participate, 13 completed round one of the study. Based on the interitem reliability analysis and the qualitative data obtained from the teachers, the decision was made to revise three of the rubrics. These rubrics are: Interactivity, How Scientists work, and Inquiry.

Interim Phase

Following the revision of the rubrics, participants were asked to reevaluate the Curve Ball website using only the revised rubrics. Of the 13 teachers who responded during Phase One, six responded to the request for feedback during the Interim phase. The data collected during this phase indicated that the changes to the rubrics provided better definitions of the overarching concept and more clarity of distinction between the levels of quality.

Phase 2

During Phase Two, the teachers were asked to use the revised rubrics to evaluate the Bucket Buddies website. Of the 12 participants evaluating the Curve Ball website in Round One, 12 participants completed round two. Both the quantitative data and the qualitative data suggest that the revisions to the rubrics strengthened the reliability of the instrument.

Data analysis

It is important to note that the number of items in this scale (8) and the number of participants in the study (13) are small, which will affect the alpha reliability coefficient. As Pedhazer states "the reliability of the instrument will increase as a result of adding items that measure the same phenomenon" (Pedhazer, p. 101). Therefore, it is important to draw on more sources of information than simply running a reliability analysis of the instrument. For this reason, both quantitative and qualitative data were collected and analyzed to determine which rubrics needed to be revised.

The following quantitative analyses were used:

- Descriptive Statistics
- Inter-Item Reliability

The following qualitative methods were used:

- Written comments were collected when the participants evaluated the website
- Follow-up comments were collected to provide additional information related to revising the rubrics

Quantitative Analysis

Descriptive Statistics

Descriptive statistics were run for the eight rubrics for both Phase One and Phase Two data. Table One shows the descriptive statistics for Phase One data and Table Two shows the descriptive statistics for Phase Two data.

For the Phase One data, the descriptive statistics reveal that one of the rubrics has a standard deviation of .000 indicating that the raters were in perfect agreement. This is problematic for the reliability analysis, as will be discussed shortly. Of the other seven rubrics, three rubrics had data at only two of the three levels indicating that the raters were in close agreement. Of the four remaining rubrics, "Work" and "Inquiry" have large standard deviations, indicating that there was a substantial amount of disagreement among raters. See Table 1.

Table 1: Descriptive Statistics for Phase 1

	N	Minimum	Maximum	Mean	Std. Deviation
AUTHORITY	13	3	3	3.00	.000
DESIGN	13	2	3	2.85	.376
INTERACTIVITY	13	2	3	2.85	.376
COMMUNICATION	13	1	3	2.08	.494
INQUIRY	13	1	3	1.69	.751
WORK	13	1	3	1.69	.751
WRITING	13	2	3	2.92	.277
INTEGRATION	13	1	3	2.54	.660

For the Phase Two data, the descriptive statistics reveal that one of the rubrics has a standard deviation of .000 indicating that the raters were in perfect agreement. Again, this is problematic for the reliability analysis, as will be discussed shortly. Of the other seven rubrics, three rubrics had data at only two of the three levels indicating that the raters were in close agreement. Of the four remaining rubrics, "Interactivity", "Inquiry" and "Work" have large standard deviations, indicating that there was a substantial amount of disagreement among raters. See Table 2.

Table 2: Descriptive Statistics for Phase 2

	N	Minimum	Maximum	Mean	Std. Deviation
AUTHORITY	12	2	3	3.92	.289
DESIGN	12	2	3	2.92	.289
INTERACTIVITY	12	1	3	1.50	.800
COLLABORATION	12	3	3	3.00	.000
INQUIRY	12	1	3	2.50	.800
WORK	12	1	3	1.58	.800
WRITING	12	2	3	2.92	.289
INTEGRATION	12	1	3	2.75	.621

It is interesting to note that in both Phase One and Phase Two, "Inquiry" and "Work" have large standard deviations. While participants indicated in their comments that they felt the distinctions between levels of quality were better, the quantitative data does not support these comments.

Counts

An examination of the data (see Table 3: Counts), shows the magnitude of the disagreement among raters for the "Inquiry" and "Work" rubrics. In both cases, two raters ranked the website as "Good", five raters ranked the website as "Fair", and six raters ranked the websites as "Poor".

Table 3: Phase One Data, Counts			
	Good	Fair	Poor
AUTHORITY	13	0	0
DESIGN	11	2	0
INTERACTIVITY	11	2	0
COMMUNICATION	2	10	1
INQUIRY	2	5	6
WORK	2	5	6
WRITING	12	1	0
INTEGRATION	8	4	1

In contrast, the data for the “Inquiry” and “Work” rubrics show closer inter-rater agreement in the Phase Two data. For both rubrics, a solid majority was in agreement. For the “Inquiry” rubric, eight raters ranked the website as “Good”, two raters ranked the website as “Fair”, and two raters ranked the websites as “Poor”. For the “Work” rubric, two raters ranked the website as “Good”, three raters ranked the website as “Fair”, and seven raters ranked the websites as “Poor”.

Table 4: Phase Two Data, Counts			
	Good	Fair	Poor
AUTHORITY	11	1	0
DESIGN	11	1	0
INTERACTIVITY	2	2	8
COMMUNICATION	12	0	0
INQUIRY	8	2	2
WORK	2	3	7
WRITING	11	1	0
INTEGRATION	10	1	1

A third rubric, “Interactivity” also had a large standard deviation. Inter-rater agreement on this rubric was good during Phase One, but slipped somewhat during Phase Two. However, an examinations of the counts reveals that the inter-rater agreement is still quite good; eight raters ranked the website as “Good”, two raters ranked the website as “Fair”, and two raters ranked the websites as “Poor”.

Reliability Analysis

Of interest to this study is the reliability of the scale, taken both as a whole as well as reliability of individual rubrics that comprise the scale. While it is not possible to obtain a reliability coefficient for a single item scale, it is possible to infer a single item’s reliability based on its relationship to other items in the scale. Furthermore, it is not considered wise to use a single item scale as; “Single item measures are generally deficient both with respect to validity and reliability” (Pedhauzer, Measurement, Design and Analysis, p. 122).

While the rubrics were designed to stand alone, the decision was made to combine the rubrics into one scale and to use inter-item reliability analysis to determine the alpha reliability coefficient for the instrument. According to Pedhauzer:

> An instrument that is internally consistent is not necessary homogeneous (eg. measuring one construct) [and] alpha should not be taken as an index of the homogeneity of an instrument. …Even when relations among items of a measure tend to be low, indicating that, by and large, they appear to be measuring different things, the total variance will become increasingly larger, relative to the sum of the variances of the items, as the number of such items is increased. Stated differently, given a sufficiently large number of items, a measure may be shown to have high internal-consistency reliability, even when it is composed of items that share little among themselves" (Pedhazur, p. 101-102).

In running the reliability analysis of the scale, one problem became immediately evident. The rubric "Authority" has a standard deviation of .000 in the Round One data, and "Communication" has a standard deviation of .000 in the round Two data. SPSS only includes items with a standard deviation grater than .00 in an interitem reliability analysis. Therefore, for both Round One and Round Two, the analysis was conducted on seven-item, rather than eight-item scales.

For Round One data, on the seven-item scale (lacking the "Authority" rubric), the reliability analysis revealed that the scale has an alpha reliability coefficient of .4937. This is not very high. However, scales with a larger number of items tend to have higher reliability coefficients, and because an eight-item scale is not large to begin with, the loss of an item may have served to decrease the Alpha reliability coefficient of the scale. Also, the number of participants is small, which may also contribute to a lower reliability coefficient.

Table 5: Phase One, Reliability Analysis

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
AUTHORITY	N/A	N/A	N/A	N/A	N/A
DESIGN	13.7692	3.3590	.1863	.2597	.4763
INTERACTIVITY	13.7692	3.3590	.1863	.2352	.4763
COMMUNICATION	14.5385	2.7692	.4527	.5056	.3667
INQUIRY	14.9231	2.9103	.1101	.2500	.5392
WORK	14.9231	2.0769	.5152	.3658	.2741
INTEGRATION	14.0769	2.5769	.3508	.6381	.3940
WRITING	13.6923	3.8974	-.1990	.1552	.5566
Alpha = .4937					

Based on an analysis of the data collected during Phase One, three rubrics were modified to strengthen their reliability. The Alpha reliability coefficient for Phase Two is .5512, in contrast, the Alpha reliability coefficient for Phase One is only .4937. Therefore, the data collected during Phase Two of the study indicates that the modifications to the rubrics did increase the reliability of the instrument.

Table 6: Phase Two, Reliability Analysis

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
AUTHORITY	14.1667	4.8788	-.1188	.5193	.5963
DESIGN	14.1667	5.0606	-.2566	.7045	.6180
INTERACTIVITY	15.5833	3.7197	.1477	.7143	.5866
COMMUNICATION	N/A	N/A	N/A	N/A	N/A
INQUIRY	14.5833	2.6288	.5974	.8839	.3320
WORK	15.5000	3.0000	.4302	.8120	.4364
INTEGRATION	14.1667	4.1515	.4894	.8879	.4905
WRITING	14.3333	3.1515	.5767	.9550	.3808
Alpha = .5512					

Qualitative Analysis

During all phases of data collection, the raters were asked to explain why they chose a particular level of quality. It was hoped that an analysis of their comments would reveal information important to the rubric revision process. For the sake of brevity, an analysis of comments is included only for those rubrics that were determined to be problematic in the quantitative analysis of the data. Appendix A shows the recommendations for modifications to the rubrics based on Phase One data.

Rubric 3: Interactivity

Phase One

During Phase One, the raters were in close agreement; of the 13 raters responding, 11 gave the site a score of Good", 2 raters ranked the site as "Fair". The comments of one rater who rated the site as "fair" reveal that his/her interpretation of interactivity was interactive communication, rather than interactivity with the website:

> *"Interactivity is fair, with links on each page for e-mail to site officials. The index foilindex.html page does link to videoconferencing with the UK, which starts on a trail of other links leading toward live real-time interactive communication, but this is buried, and can not be highlighted as a strength of this site."*

The other rater who considered the site "fair" did so because "with the lesson outlines a specific task for the student to complete that does not seem to take full advantage of the applet and its explanatory links." Thus, their interpretation of "Interactivity" was not just that the interaction was present but that the lesson took full advantage of it.

Recommendation: The comments of the raters who ranked the site as “Fair” indicate that they may not have completely understood the criteria for ranking. Therefore, the recommendation was made to more clearly defined in the criteria for ranking.

Phase Two

During Phase Two, the raters were in close agreement, but not in as close agreement as Phase One; of the 12 raters responding, two gave the site a score of Good”, two raters ranked the site as “Fair”, and eight ranked the website as “Poor”.

The comments of the raters who ranked the website as “Fair” or “Good” reveal that they did so because of the presence of a link to another website that contained interactivity.

> *I found one page (linked) where the student can ID various critters found in a pond. There were three pictures offered and online you ID one of them. There was feedback on this right away for the student.*
>
> *Page allows minimum degree of interactivity- learner is able to select invertebrates to get information about them by following links on Resource page*

Recommendation: The rubric may need to be revised to specifically indicate that linked activities are either valid or invalid for evaluation.

Rubric 5: Scientific Inquiry Ranking

Phase One

Phase One data indicated that the raters were not in close agreement; of the 13 raters responding, two gave the site a score of “Full/Complete Inquiry”, five gave it a score of “Partial Inquiry” and six gave it a score of “Initial/Preliminary Inquiry”.

Those raters giving the site a score of “Initial/Preliminary” felt that the website contained limited opportunities for inquiry as most of the activity was directed by the information provided.

> *The site outlines a procedure for the investigation and provides a data collection worksheet. The student is instructed to “identify … explain … and demonstrate explanation” which are tasks that relate to an initial stage of inquiry. The exercise does not call for higher-level thinking, make any explicit connections to real-world applications, or call upon students to form any what-if?*
>
> *Data collection is available but prescribed; extension provided in step 6 could lead to open inquiry but is limited, minimal and not implicitly assigned.*

Their ranking is a result of the rubric criteria for “Preliminary” that stated:

> *Provides a specific step-by-step procedure for the investigation and data collection techniques-prescriptive in nature as students record data as they conduct their inquiry (either hands-on or online) Data collection tables or examples are provided.*

In contrast, raters who ranked this site as a “Partial Inquiry” made comments related to higher levels of inquiry. For example:

> *The site guides the students to use the table available on the website or use one that’s printable. It leaves the door open for students to use their own data collection and the questions given within the directions are left vague to allow for individual differentiation.*
>
> *Site encourages students to reflect on their observations. Students are asked to show explain their results and show all supporting data.*

Recommendation: The data indicated that the level “Partial” is open to interpretation and that the rubric needs more distinction between the “Preliminary” and “Partial” levels of Scientific Inquiry. Therefore, the recommendation was made to make more distinct the Preliminary and Partial levels of the rubric.

Phase Two

Phase Two data indicates that while still not in close agreement, the raters are in closer agreement than in Phase One of data collection; of the 12 raters responding, eight gave the site a score of “Full/Complete Inquiry”, two gave it a score of “Partial Inquiry” and two gave it a score of “Initial/Preliminary Inquiry”.

One teacher acknowledge that deciding between “Partial” and “Full” was difficult because the website did not contain all of the criteria necessary to make it “Good”.

> *"Tricky. I would say that this site is good (full/complete inquiry) because it fulfills one or two of the criteria in the good section.*

The primary difference in responses between those teacher who gave the website a rating of “Full” and those who ranked it lower was the mention of the development of a hypothesis to begin the activity.

> *Students begin with development of a hypothesis. Students use scientific tools to collect data and provide scientific data leading to the ability to think critically and logically about the effects of organisms in a habitat.*

One teacher who ranked the website as “Fair” acknowledged the presence of developing a hypothesis but had the following comment.

> *With the teacher they come up with a hypothesis. They then investigate this question as a class and compare their data with other schools…*

Both teachers who ranked the site as “Preliminary” did not mention the development of a hypothesis but indicated that the activity was prescriptive in nature:

> *Provides a procedure for the investigation and data collection techniques-prescriptive in nature as students record data as they conduct their inquiry (either hands-on or online).*
>
> *Step by step directions inform students how to collect samples and identify them.*

It may be that the teachers who rated the site as Preliminary Inquiry, simply missed the part of the activity dealing with developing a hypothesis.

Recommendation: Base on the available data, there is no need to revise the rubric. However, interviews with the two teachers who ranked the site as “Preliminary” may provide additional information for strengthen the definitions of the rubric.

Rubric 6: How Scientists Work

Phase One

During Phase One of data collection, the raters are not in close agreement; of the 13 raters responding, two gave the site a score of “Good”, five gave it a score of “Fair”, and six gave it a score of “Poor”.

The raters who ranked this site as poor felt that the site did not provide examples of how scientists work.

> *This page provides no insight as to how scientists work.*
>
> *I did not find any reference or links to help students understand the nature of inquiry or how scientists work and it is not stated on the site itself.*
>
> *Although it demonstrates one of the methods of scientific exploration, there is no “Opportunity to Understand” why one does this.*

In contrast, the raters who ranked the site as “Fair” thought that the opportunity to see how scientists work was provided by the activity, which allows students to act like a scientist.

> *Site offers students the opportunity to explore several different kinds of science investigations.*

> *This site is focused on holding the student's hand through a specific, very complex analysis, and reporting of results. This provides first hand experience in scientific analysis. Some other aspects of scientific work, like hypothesis testing and maintaining a skeptical attitude, are less emphasized.*

One rater who ranked the site as "Fair" mentioned the existence of a model as a method of conveying scientific information.

> *Presents model of pitch and explains the necessity of a model and that it is a simplified version of reality – points out that it is really much more complicated and why.*

One rater who ranked the site as "Good" also used the inquiry activity as justification for the ranking.

> *The activity presents students with a real life problem that gives them insight into scientific procedures used by scientists. Students gather data and propose a problem leading to the formation of a solution. Connections are made between the activity and real world problems.*

Recommendation: Phase One data indicated that the rubric needed to be revised to specifically indicate whether or not the presence of an inquiry activity is sufficient to rank the site as "Fair" or "Good" on the criteria of "How Scientist Work" The rubric was revised to clearly indicate that the presence of inquiry activities is not a criteria for ranking a site as "Fair" or "Good".

Phase Two

During Phase Two of data collection, the raters are in closer agreement than during Phase One; of the 12 raters responding, two gave the site a score of "Good", three gave it a score of "Fair", and seven gave it a score of "Poor".

Some of the comments gathered indicate that there may still be some confusion about the definition of what comprised "How Scientists Work:

> *"Students simulate scientific investigative activities including the use of microscopes and insect-gathering materials. Students also collect data, organize information and report information to compare with other student 'scientists.'"*

The rubric specifically states that the presence of inquiry activities is not a criteria for evaluation a website as "Far" or "Good" so this is a training issue and does not warrant a change to the rubric.

However, there is a comment that may warrant a change to the rubric.

> *"While the project is well structured and implicitly reflects the ways the scientists work, the material explicitly provided is accurate but brief. There is a link to USGS Water Science for Schools web site that provides more specific information."*

Recommendation: Like the Interactivity rubric, the criteria may need to be defined to specifically state whether a link out to information ***at another website*** is sufficient to give the website in question a higher ranking.

Rubric 8: Resource Integration Ranking

Phase One

In Phase One of data collection, the raters are in agreement; of the 13 raters responding, seven raters ranked the site as "Good", four ranked the site as "Fair", and one ranked the site as "Poor".

The comment of rater who ranked the site as Poor, indicate that they did so because answers were not readily available on the website. This however, relates more to the criteria of Interactivity than the criteria of Resource Integration:

> *The "answer" link states simply that "answers will vary". That's not very satisfactory, particularly if a student is not sure of a concept. Suggested responses within given parameters would be helpful.*

Raters who ranked the website as "Fair" mentioned the downloadable worksheet. Since the criteria for "Fair" is the presence of among other things, a downloadable worksheet, this criteria is met by the website.

> *There is a printable worksheet for students to use. No timeline for how long the lesson should take, or how to integrate it into existing lessons on aerodynamics. It appears to be a stand-alone activity.*
>
> *From this lesson page there are links to download the applet and worksheet and a link to the index of the Glenn Research Center that provides additional information.*

The majority of the raters thought that the downloadable worksheet, in conjunction with the link to standards and the teacher plans warranted a ranking of "Good".

> *Supporting materials are available through links. There are activity directions and a worksheet for students to record data*
>
> *Articulation to the Standards is explicit, via links provided.*

> *By mentioning the standards involved, a teacher can see where to integrate this into their curriculum.*
>
> *Handouts available for students and referenced in teacher plans. Plans provide minimum options for classroom adaptation as far as groups, using computers or using paper.*

Recommendation: The Phase One data indicated that the rubric needed to more clearly define the levels of quality of “Resource Integration”.

Phase Two

Phase Two data indicates that inter-rate agreement is better with the revised rubric; of the 12 raters responding, 10 raters ranked the site as “Good”, one ranked the site as “Fair”, and one ranked the site as “Poor”.

Recommendations: There is no need for further revisions.

Appendix A: Recommendations Based on Phase One Data

Rubric 1: Authority

This was not a good pilot test of the Authority rubric. A less authoritative site should be used in the second round of data collection.

Rubric 2: Design

The data does not indicate a need to revise the rubric.

Rubric 3: Interactivity

The data indicates that the rubric needs to define the concept “Interactivity”.

Rubric 4: Communication

The data does not indicate a need to revise the rubric.

Rubric 5: Scientific Inquiry

The data indicates that the rubric needs more distinction between the “Preliminary” and “Partial” levels of Scientific Inquiry.

Rubric 6: How Scientists Work

Rubric needs to be revised to specifically indicate whether or not the presence of an inquiry activity is sufficient to rank the site as “Fair” or “Good” on the criteria of “How Scientist Work”

Rubric 7: Quality of Writing

The data does not indicate a need to revise the rubric.

Rubric 8: Resource Integration

The data indicates that the rubric needs more clearly defined levels of “Resource Integration”.

Appendix B: Recommendations Based on Phase Two Data

Rubric 1: Authority

The data does not indicate a need to revise the rubric.

Rubric 2: Design

The data does not indicate a need to revise the rubric.

Rubric 3: Interactivity

The criteria may need to be defined to specifically state whether a link out to information ***at another website*** is sufficient to give the website in question a higher ranking.

Rubric 4: Communication

The data does not indicate a need to revise the rubric.

Rubric 5: Scientific Inquiry

Base on the available data, there is no need to revise the rubric. However, interviews with the two teachers who ranked the site as "Preliminary" may provide additional information for strengthen the definitions of the rubric.

Rubric 6: How Scientists Work

The criteria may need to be defined to specifically state whether a link out to information ***at another website*** is sufficient to give the website in question a higher ranking.

Rubric 7: Quality of Writing

The data does not indicate a need to revise the rubric.

Rubric 8: Resource Integration

The data does not indicate a need to revise the rubric.