International Journal on New Computer Architectures and Their Applications (IJNCAA) 1(3): 665-682
The Society of Digital Information and Wireless Communications, 2011 (ISSN: 2220-9085)

Blog Quality Measurement: Analysis of Criteria using The Rasch Model

Zuhaira Muhammad Zain[1] , Abdul Azim Abd Ghani[1], Rusli Abdullah[1], Rodziah Atan[1], and Razali Yaakob[2]
1 Department of Information Science, Universiti Putra Malaysia, Serdang, Selangor
2 Department of Computer Science, Universiti Putra Malaysia, Serdang, Selangor
Mailing Address
zuhaira.muhdzain@gmail.com, {azim, rusli, rodziah, razaliy}@fsktm.upm.edu.my

ABSTRACT

Research in blog quality has become increasingly important, as preferences evolve in the way people gain information. For this study, blog quality categories and criteria were derived from metadata analysis and recent literature and then tested in two surveys. Rasch model analysis of responses provides systematic evidence of construct validity for the 11 quality categories and 49 criteria. The first survey, addressed to expert reviewers, supports the content aspect of construct validity, with one modification to a quality category. The second survey, given to blog readers, finds strong agreement with the quality items after the removal of three criteria because of redundancy. The second survey supports the substantive, structural, generalizability, external and consequential aspects of construct validity. These results constitute an important step toward development of a valid and widely applicable blog quality model.

KEYWORDS

blog; blog quality; blogging; blogosphere; construct validity; quality; quality assessment; quality measurement; questionnaire quality; Rasch model; rating scale; survey reliability.

1 INTRODUCTION

This study examines proposed criteria for blog quality, with a view to confirming that these criteria will promote reader satisfaction. The criteria are derived from a thorough analysis of metadata and from research of the literature in related areas, including blogging [1,2,3], website design [4], information quality on the Web [5,6,7,8,9,10] and portal data quality [11]. Prior to this study, these criteria have not been reviewed and their validity has not been verified systematically.

Providing empirical evidence for validity is a basic requirement in development of a reliable survey instrument for assessment of blog quality. The purpose of this study is to test the criteria for several aspects of construct validity, as proposed by Messick [12], including content, substantive, structural, generalizability, external and consequential aspects of construct validity. In order to achieve this objective, two tests are conducted: a content validity test and a pilot test.

The content validity test is an assessment of items in a survey instrument by a group of expert reviewers. It involves a systematic review of the survey's contents to ensure

that it includes everything it should and excludes everything that should not be included. This step is important in providing a good foundation on which to base a rigorous assessment of validity. Although Kitchenham and Pfleeger [13] claim that there is no content validity statistic, this argument has been refuted by Abdul Aziz et al. [14], who confirm that the content aspect of construct validity can be assessed accurately by using the Rasch measurement model. The Rasch model takes account of both the ability of respondents and the difficulty of questionnaire items [15]. The graphical output provided by this technique facilitates quick and easy decision making [16]. In this study, the Rasch model is applied to the content validity test to confirm whether the quality categories and the quality criteria in each category enjoy consensus among the reviewers. The results provide empirical evidence to support the content aspect of construct validity for a blog quality criteria survey.

The pilot test addresses the other five aspects of construct validity. Fisher [17] asserts that the Rasch model is a tool of construct validation. Bond [18] and Wolfe and Smith [19] provide guidelines on how Rasch analyses help in eliciting evidence to support Messick's unified validity. In order for the survey instrument to be applied reliably to other samples with replicable results, it should show a reasonable-level-of-accuracy value within the confidence interval. If the generated accuracy value is not acceptable, the instrument has to undergo amendments until it is able to show reliability within the confidence interval. Correct measurement leads to correct analysis and correct assessment.

The results from the content validity test, administered by means of an online survey, provide empirical evidence for the content aspect of construct validity for the 11 quality categories and the 49 quality criteria. Results from the pilot test support substantive, structural, generalizability, external and consequential aspects of validity for the 49 criteria. The pilot test was administered by manual distribution.

A valid model of blog quality can benefit bloggers to determine which criteria to focus when designing their blog. It also has a potential use as a valid guideline for blog readers or evaluators to check whether the visited blog is of quality or not. It is also crucial to keep only the good quality blog in the blogosphere.

The rest of this paper is organized as follows: Section 2 describes the basics of the Rasch measurement method; Section 3 explains how the content validity and pilot tests were conducted; Section 4 discusses the results; finally, Section 5 touches on conclusions and future work.

2 RASCH MEASUREMENT METHOD

The Rasch model offers a mathematical framework against which researchers can compare their data. Its basic idea is that a useful measurement entails assessment of only one item at a time (unidimesionality) on a hierarchical line of inquiry [20]. By using the theoretical idealization, patterns of responses that do not match with this ideal can be compared. Furthermore, person and item performance that deviate from that line (fit) can be measured. Therefore, the item wording and score interpretations from these data can be reconsidered by the researcher.

In this study, responses from the expert reviewers (content validity test)

and responses from blog readers (pilot test) are considered as a rating scale. The respondents rated the blog quality criteria according to their level of agreement with each item. In this phase, the study is only counting the number of positive answers, which are added up to give a total raw score. The raw score provides a ranking order, which serves as an ordinal scale reflecting a continuum of response [21]. The data are not divided into equal intervals, which contradicts the way numbers are used in statistics, and they do not meet the fundamental requirements for statistical evaluation [22]. Rasch analysis can solve this problem by providing a transformation of an ordinal score into a linear, interval-level variable by estimating fit of the data to the Rasch model expectations.

Rather than fitting collected data to a measurement model with errors, the Rasch model focuses on perfecting the survey instrument, so that it measures with accuracy. By emphasizing the reproducibility of the latent trait measurement, this approach gives reliability its rightful place in supporting validity. Measuring blog quality criteria in an appropriate way is vital to ensuring valid quality information. The Rasch method absorbs error to presenting a more accurate prediction based on a probabilistic model [23].

In the Rasch measurement model, when an individual respondent's level of ability has been determined (in our case, the level of agreement by expert reviewers and blog readers, represented as β_v) and the item difficulty has been estimated (in our case, the level of agreement to an item, or δ_i), then the probability of a successful response (in our case, a blog quality criterion being affirmed) can be expressed mathematically as

$$P(\theta) = \frac{e^{\beta_v - \delta_i}}{1 + e^{\beta_v - \delta_i}} \quad (1)$$

where

e = base of natural logarithm or Euler's number: 2.7183
β_v = person's ability
δ_i = item or task difficulty

3 METHODOLOGY

3.1 Content Validity Test

The pool of expert reviewers in the content validity test comprised 50 university lecturers in English from various institutions in Malaysia and 50 information technology executives or managers with more than 10 years' experience.

The study design was based on the objective of gathering evidence about the validity of blog quality criteria. A questionnaire was developed to determine whether the experts agreed with the proposed set of quality categories and the assignment of quality criteria to their respective categories. The questionnaire asked for Yes/No answers but also provided space for elaboration of differing views and comments. The experts' opinions were of interest for potential modifications to the blog quality instrument.

E-mail invitations to join the online survey covered the objective of the study, its relevance, the importance of the individual's participation and an assurance of confidentiality. Responses were tabulated and analysed using the basic Rasch dichotomous model [24].

3.2 Pilot Test

Forty blog readers from the faculty of the Computer Science & Information Technology department, Universiti Putra Malaysia, participated in the pilot test. Their questionnaires asked them to state their level of agreement with each of the 49 blog quality criteria on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). Data were tabulated and analysed using the Rasch rating scale model [25].

4 RESULTS AND DISCUSSION

4.1 Content Validity Test

Figure 1 shows the summary statistics for the analysis of the sample of 60 experts (survey response = 60%) on the 63 dichotomous scale items that comprise the content validity test for blog quality categories and criteria. The mean of the individual person measures is 2.77 (SE .56), which is noticeably higher than the 0 calibration of the quality item scale, which is set as the default option of the analysis. The standard deviation of the person measures is 1.18 logits, while the standard deviation for quality item measures diverges even further to 1.33. The summary fit statistics for quality items and persons show satisfactory fit to the model. The quality item reliability is similar to the person reliability (.74). This indicates that the survey instrument for measuring content validity is reliable and results are reproducible.

Persons	60 INPUT	60 MEASURED		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	49.3	57.0	2.77	.56	1.01	.1	1.28	.2
S.D.	5.4	.0	1.18	.23	.16	.6	1.64	1.0
REAL RMSE	.60	ADJ.SD	1.02	SEPARATION	1.70	Person RELIABILITY		.74
Items	63 INPUT	63 MEASURED		INFIT		OUTFIT	
MEAN	41.5	48.0	.00	.63	1.00	.1	1.25	.3
S.D.	6.8	.0	1.33	.25	.11	.4	1.60	.9
REAL RMSE	.68	ADJ.SD	1.14	SEPARATION	1.67	Item RELIABILITY		.74

Figure 1. Summary statistics

The Wright map in Figure 2 displays the distribution of experts on the left and the distribution of item agreement on the right according to item number. The most agreed-to items are items 55 (*Availability of blog*), 51 (*Easy to read information*), 50 (*Readability*), 34 (*Information Representation*), 17 (*Currency*) and 9 (*Appropriate explanatory text*). These are located at -2.91 logits (SE 1.85). The least agreed-to item is item 40 (*Must-have sound*), located at the top of the item distribution at +3.51 logits (SE .34). The person distribution confirms the result from the summary statistics. The most agreeable experts are r11, r28, r39, r41, r52 and r58, and these are located at +4.98 logits (SE 1.04). The least agreeable expert is r29, located at the bottom of the person distribution at +.07 logits (SE .31). We noted earlier in this section that the mean of the person distribution is higher than the mean of the item distribution. This indicates that all experts involved in the content validity test tend to agree to the entire set of quality categories and their assigned criteria. The probability of agreement by the experts to the quality

categories and criteria can be established by using formula (1):

$$P(\theta) = \frac{e^{2.77-0}}{1+e^{2.77-0}}$$

$$= 0.941$$

Thus, the expert reviewers in the content validity test indicate their level of agreement to the quality categories and criteria at 94.1%, which is above the 70% limit of Cronbach's alpha. Hence, all experts agree to the proposed quality categories and their assigned criteria.

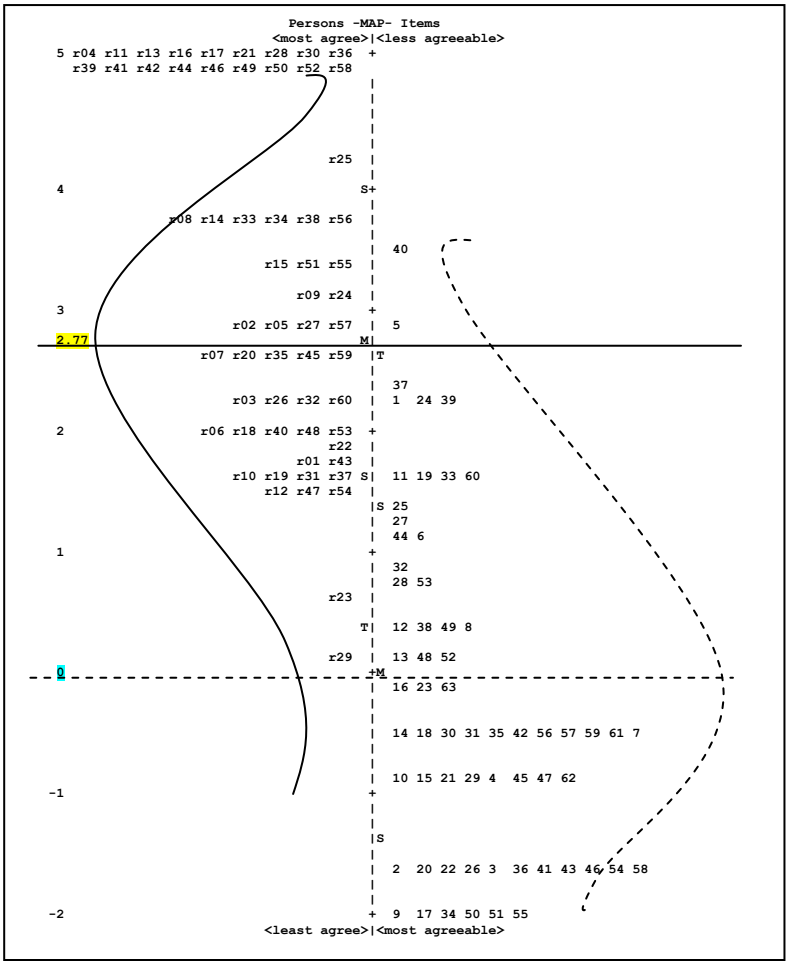

Figure 2. Wright map

Figure 3 displays the quality-item statistics in measure order. Thus, the topmost item and bottommost items on both the Wright map and the table of item statistics correspond. For any form of genuine scientific investigation, unidimensionality is a requirement. Inspection of the Rasch fit statistics for quality items is the first step towards examining the dimensionality of this test. The fit statistics reveal that there are six minimum-estimated items which are 100% agreed to by the experts. These correspond to the most agreed-to items on the Wright map. They are kept in this analysis because they cause no threat to measurement. Checking the Outfit MNSQ and Outfit Z-Std columns, we

find that while nearly all of the items sufficiently fit to the model, there are two misfits. Their Outfit MNSQ value > 1.4 and Outfit Z-Std value > +2.0 indicate that they behaved more erratically than expected.

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	Item
40	17	48	3.51	.34	.95	-.2	1.40	1.6	40-Must-have sound
5	23	48	2.84	.33	.93	-.5	.94	-.3	5-Relevant info
37	27	48	2.41	.33	.99	.0	1.10	.5	37-Info in diff format
24	28	48	2.30	.33	.90	-.7	.78	-1.0	24-Easy to remember add
39	28	48	2.30	.33	.91	-.7	.81	-.8	39-Must-have photos
1	29	48	2.19	.33	1.27	2.0	1.48	1.8	1-F.Accuracy
11	34	48	1.62	.35	.89	-.7	.76	-.7	11-Avail. Blog owner info
19	34	48	1.62	.35	.86	-.9	.68	-1.0	19-Real-occurrences info
33	34	48	1.62	.35	1.03	.3	1.01	.1	33-Technorati rank
60	34	48	1.62	.35	1.17	1.1	1.08	.3	60-Chat box
25	36	48	1.37	.36	.97	-.1	1.54	1.4	25-Emotional support
27	37	48	1.23	.37	1.15	.8	.91	-.1	27-Personal feel
6	38	48	1.08	.39	1.07	.4	1.18	.5	6-Originality
44	38	48	1.08	.39	1.01	.1	.97	.1	44-Interactivity
32	39	48	.93	.40	.91	-.4	.69	-.6	32-Rewarding experience
28	40	48	.76	.42	1.00	.1	.82	-.2	28-Surprises
53	40	48	.76	.42	1.06	.3	.93	.0	53-Readable font
8	42	48	.38	.46	1.03	.2	1.10	.4	8-Amount of info
12	42	48	.38	.46	1.04	.2	.79	-.1	12-Easy to understand
38	42	48	.38	.46	1.02	.2	.86	.0	38-Multimedia
49	42	48	.38	.46	1.08	.4	1.37	.7	49-Intuitive interface
13	43	48	.15	.50	.97	.0	.59	-.4	13-Informative
48	43	48	.15	.50	.98	.1	1.21	.5	48-Good use of colours
52	43	48	.15	.50	.93	-.1	.66	-.3	52-Legibility
16	44	48	-.13	.55	1.13	.4	.78	.0	16-Link to info
23	44	48	-.13	.55	1.10	.4	.77	.0	23- Cognitive advancement
63	44	48	-.13	.55	.75	-.5	.37	-.7	63-Trackback
7	45	48	-.46	.62	1.18	.5	1.30	.6	7- F.Completeness
14	45	48	-.46	.62	.81	-.3	1.10	.4	14-Objective info
18	45	48	-.46	.62	1.11	.4	.85	.2	18-Real time info
30	45	48	-.46	.62	1.10	.4	.74	.1	30-Reputation of blog
31	45	48	-.46	.62	.90	.0	.79	.1	31- Reputation of blogger
35	45	48	-.46	.62	.97	.1	.54	-.2	35-Exciting content
42	45	48	-.46	.62	1.00	.2	.84	.2	42-Ease of ordering
56	45	48	-.46	.62	.87	-.1	.48	-.3	56-Blog responsiveness
57	45	48	-.46	.62	.87	-.1	.48	-.3	57-Ease of info access
59	45	48	-.46	.62	.91	.0	.82	.2	59-Blogroll
61	45	48	-.46	.62	.93	.0	.43	-.4	61-Comment field
4	46	48	-.92	.75	.91	.1	7.69	2.8	4-Reliable source
10	46	48	-.92	.75	1.13	.4	2.48	1.3	10-Appropriate level
15	46	48	-.92	.75	1.14	.4	1.34	.7	15-Provide info sources
21	46	48	-.92	.75	1.08	.3	.72	.1	21-F.Engaging
29	46	48	-.92	.75	1.12	.4	1.44	.7	29-F.Reputation
45	46	48	-.92	.75	1.08	.3	.76	.2	45-F.Visual design
47	46	48	-.92	.75	1.05	.3	.57	.0	47-Clear layout
62	46	48	-.92	.75	.83	-.1	.36	-.3	62-Search tool
2	47	48	-1.67	1.03	.80	.1	.14	-.6	2-Correct info
3	47	48	-1.67	1.03	1.11	.4	4.61	1.9	3-Reliable info
20	47	48	-1.67	1.03	.80	.1	.14	-.6	20-Up-to-date
22	47	48	-1.67	1.03	1.09	.4	1.52	.8	22- Appreciate comments
26	47	48	-1.67	1.03	1.12	.4	9.90	3.8	26-Fun
36	47	48	-1.67	1.03	.91	.2	.21	-.4	36-Fresh perspective
41	47	48	-1.67	1.03	1.08	.4	1.04	.5	41-F.Navigation
43	47	48	-1.67	1.03	1.10	.4	2.41	1.2	43-Easy to navigate
46	47	48	-1.67	1.03	.91	.2	.21	-.4	46-Attractive layout
54	47	48	-1.67	1.03	1.09	.4	1.52	.8	54-F.Blog accessibility
58	47	48	-1.67	1.03	.80	.1	.14	-.6	58-F.Blog Tech Features
9	48	48	-2.91	1.85	MIN ESTIMATED MEASURE				9-Appropriate exp. text
17	48	48	-2.91	1.85	MIN ESTIMATED MEASURE				17-F.Currency
34	48	48	-2.91	1.85	MIN ESTIMATED MEASURE				34-F.Info representation
50	48	48	-2.91	1.85	MIN ESTIMATED MEASURE				50-F.Readability
51	48	48	-2.91	1.85	MIN ESTIMATED MEASURE				51-Easy to read info
55	48	48	-2.91	1.85	MIN ESTIMATED MEASURE				55-Availability of blog

Figure 3. Item Measure – Acceptable range for Infit and Outfit Mean-square is between 0.6 to 1.4 [26] and acceptable range for Infit and Outfit Z-std is between -2 to +2 [20]

Counterchecking against the Guttman scalogram (see Figure 4) indicates that the two items, *Reliable source* (item 4) and *Fun* (item 26), were underrated by respondents 41 and 58, respectively.

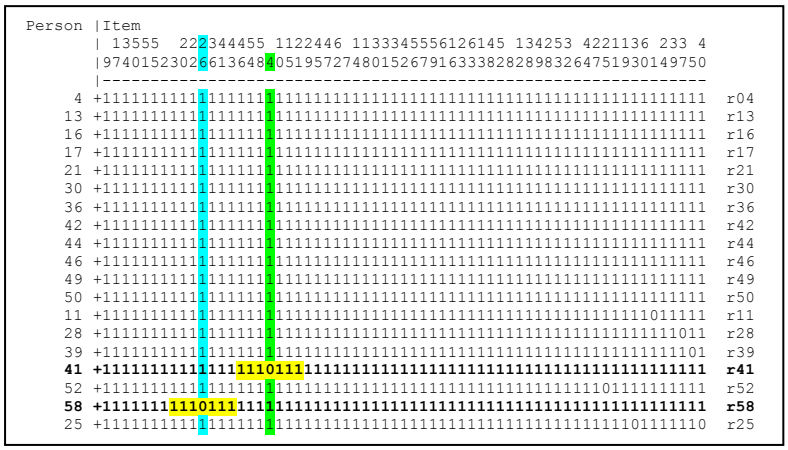

Figure 4. A segment of Guttman scalogram

One possible reason is that they were careless in completing their surveys. After verifying that the Infit values (see INFIT column in Figure 3) are within range, the two misfits are deemed acceptable. This criterion-referenced interpretation of measurements supports the technical quality of the content aspect of construct validity.

Further investigation of the person statistics (see Figure 5) confirms the result from the item statistics, where respondents 41 and 58 are considered too haphazard. However, a study of the Infit Mean-square and Infit Z-std values shows that they are within the bound; therefore, they can be accepted. The person statistics also reveal that the top 12 people with maximum estimated measures are the experts who agreed with all the category definitions and the assigned criteria. They are accepted in this analysis because they do not represent any threat to the measurement.

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
4	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
13	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
16	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
17	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
21	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
30	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
36	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
42	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
44	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
46	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
49	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
50	57	57	6.23	1.86	MAX ESTIMATED MEASURE			
11	56	57	4.98	1.04	1.04	.4	.30	-.1
28	56	57	4.98	1.04	1.01	.3	.25	-.2
39	56	57	4.98	1.04	.94	.2	.17	-.3
41	56	57	4.98	1.04	1.16	.5	6.41	2.2
52	56	57	4.98	1.04	1.11	.4	.67	.3
58	56	57	4.98	1.04	1.17	.5	9.90	3.2
25	55	57	4.20	.76	.84	-.1	.30	-.2
8	54	57	3.72	.64	1.05	.3	4.04	1.9
14	54	57	3.72	.64	1.12	.4	.97	.4
33	54	57	3.72	.64	.92	.0	.40	-.3
34	54	57	3.72	.64	1.26	.7	2.47	1.3
38	54	57	3.72	.64	.87	-.1	.39	-.3
56	54	57	3.72	.64	.89	-.1	.64	.0
15	53	57	3.36	.57	.84	-.3	.35	-.6
51	53	57	3.36	.57	1.05	.3	1.06	.4
55	53	57	3.36	.57	.83	-.3	.41	-.5
9	52	57	3.07	.52	1.15	.5	2.58	1.6
24	52	57	3.07	.52	1.21	.7	1.90	1.1
2	51	57	2.82	.48	1.32	1.0	.98	.2
5	51	57	2.82	.48	1.10	.4	.94	.2
27	51	57	2.82	.48	1.13	.5	1.61	1.0
57	51	57	2.82	.48	.85	-.4	.35	-1.0
7	50	57	2.60	.46	.95	-.1	1.61	1.0
20	50	57	2.60	.46	1.08	.4	.76	-.2
35	50	57	2.60	.46	.96	-.1	.60	-.5
45	50	57	2.60	.46	1.01	.1	1.95	1.4
59	50	57	2.60	.46	.84	-.5	.42	-.9
3	48	57	2.22	.42	1.20	.8	1.01	.2
26	48	57	2.22	.42	.90	-.3	.52	-.9
32	48	57	2.22	.42	1.23	1.0	1.82	1.4
60	48	57	2.22	.42	.61	-1.7	.32	-1.6
6	47	57	2.06	.40	.93	-.2	.83	-.2
18	47	57	2.06	.40	.99	.0	1.12	.4
40	47	57	2.06	.40	1.11	.5	.77	-.4
48	47	57	2.06	.40	.86	-.6	.60	-.8
53	47	57	2.06	.40	.74	-1.1	.55	-1.0
22	46	57	1.90	.39	1.06	.3	1.00	.2
1	45	57	1.75	.38	1.12	.6	1.09	.4
43	45	57	1.75	.38	.80	-.9	.55	-1.2
10	44	57	1.61	.37	1.00	.1	.86	-.3
19	44	57	1.61	.37	1.11	.6	1.21	.6
31	44	57	1.61	.37	.98	.0	.87	-.2
37	44	57	1.61	.37	.65	-1.8	.51	-1.5
12	43	57	1.48	.36	.98	.0	.72	-.8
47	43	57	1.48	.36	.84	-.8	.75	-.7
54	43	57	1.48	.36	1.16	.8	1.08	.4
23	35	57	.56	.32	1.10	.7	1.18	.8
29	30	57	.07	.31	1.22	1.6	1.66	2.7

Figure 5. Person measure

As stated earlier, the objective of the content validity test includes two aspects: first, the identification of quality categories and, second, the assignment of criteria to the categories. Before going on to analyse the experts' views and comments from the open question, it is necessary to calculate the percentage of probability that the two aspects would be agreed to, based on the logit measure. The purpose of this step is to decide whether the two aspects need to be reviewed. A threshold value of 70% is set, in line with the standard threshold limit of Cronbach's alpha [27]. The evaluation process can be described as follows:

- If a category definition and the assigned criteria have a probability percentage of being agreed to greater than 70%, they will be accepted without review.
- If the percentage of probability is less than 70%, they will be reviewed if

comments are provided by the experts. The category will then be redefined and its criteria will be discarded or amended as necessary.

The results for the 11 categories are presented in Table 1. For nine of the category definitions, the percentage of probability for agreement by the expert reviewers is between 70% and 95%. The *Accuracy* and *Completeness* categories need to be reviewed, as their percentages of probability are below 70%. However, the *Completeness* category is accepted without review because there are no comments available from the expert reviewers to guide redefinition. The *Accuracy* category has been redefined as suggested by reviewer comments. See Table 3 for the accepted definitions of the 11 quality categories.

Table 1. Probability percentages for agreement to each of 11 blog quality categories

Category		P(Θ) (%)	Category		P(Θ) (%)
1	Accuracy	10.07	7	Blog Accessibility	84.16
2	Completeness	61.30	8	Blog Technical Features	84.16
3	Engaging	71.50	9	Currency	94.83
4	Reputation	71.50	10	Info Representation	94.83
5	Visual Design	71.50	11	Readability	94.83
6	Navigation	84.16			

The findings for the assignment of criteria to their respective categories are shown in Table 2. It can be seen that 16 criteria (probability percentage for agreement > 70%) remain in their respective categories and 36 criteria should be reviewed. However, there are no comments available for 31 of these criteria; this means that they remain in their categories. Five criteria can be revisited: (1) *Relevant info* in the category *Accuracy*, (2) *Easy to remember address* in the category *Engaging*, (3) *Must-have sounds*, (4) *Info displayed in different format* and (5) *Must have photos*. The last three criteria are from the category *Info Representation*. As suggested by the experts, the actions taken are as follows:

- *Relevant info* is transferred to the category *Completeness*.
- *Easy to remember address* is replaced by *Memorable content*.
- *Info displayed in different format* is eliminated for having the same meaning as *Multimedia*.
- *Must-have photos* is discarded from the category *Info Representation* as it is an integral part of *Multimedia*.
- *Must-have sounds* is removed for the same reason.

Table 2. Probability percentages for agreement to quality criteria

Category		P(Θ) (%)	Category		P(Θ) (%)
1	Must-have sound	2.90	27	Objective info	61.30
2	Relevant info	5.52	28	Real time info	61.30
3	Info in different format	8.24	29	Reputation of blog	61.30
4	Easy to remember address	9.11	30	Reputation of blogger	61.30
5	Must-have photos	9.11	31	Exciting content	61.30

Table 2. Continued.

Category		P(Θ) (%)	Category		P(Θ) (%)
6	Availability of blog owner info	16.52	32	Ease of ordering	61.30
7	Real-occurrence info	16.52	33	Blog responsiveness	61.30
8	Technorati rank	16.52	34	Ease of information access	61.30
9	Chat box	16.52	35	Blogroll	61.30
10	Emotional support	20.26	36	Comment field	61.30
11	Personal feel	22.62	37	Reliable source	71.50
12	Originality	25.35	38	Appropriate level of content	71.50
13	Interactivity	25.35	39	Provide information source	71.50
14	Rewarding experience	28.29	40	Clear layout of info	71.50
15	Surprises	31.86	41	Search tool	71.50
16	Readable font	31.86	42	Correct info	84.16
17	Amount of info	40.61	43	Reliable info	84.16
18	Easy to understand	40.61	44	Up-to-date	84.16
19	Multimedia	40.61	45	Appreciate comments	84.16
20	Intuitive interface	40.61	46	Fun	84.16
21	Informative	46.26	47	Fresh perspective	84.16
22	Good use of colours	46.26	48	Easy to navigate	84.16
23	Legibility	46.26	49	Attractive layout	84.16
24	Link to info	53.25	50	Appropriate explanatory text	94.83
25	Cognitive advancement	53.25	51	Easy to read info	94.83
26	Trackback	53.25	52	Availability of blog	94.83

See Table 3 for the final assignment of the 49 criteria to the 11 quality categories. These were used in the pilot test for measuring the acceptability of criteria for blog quality.

Table 3. Final result of content validity test.

Category		Definition	Quality criteria	
1	Accuracy	The extent to which information is exact and correct, certified as being free-of-error	1	Correct information
			2	Reliable info
			3	Reliable source
			4	Originality
2	Completeness/ Comprehensiveness of Info	The extent to which the information provided is sufficient	5	Amount of information
			6	Appropriate explanatory text
			7	Appropriate level of content
			8	Availability of blog owner information
			9	Easy to understand information
			10	Informative
			11	Links to information
			12	Objective information
			13	Providing information sources
			14	Relevant info
3	Currency, Timeliness, Update	The extent to which the blog provides non-obsolete information	15	Real time info
			16	Real-occurrence info
			17	Up-to-date info

Table 3. Continued.

Category		Definition	Quality criteria	
4	Engaging	The extent to which the blog can attract and retain readers	18	Appreciation for readers' comments
			19	Cognitive advancement
			20	Emotional support
			21	Fun
			22	Surprises
			23	Personal feel
			24	Memorable content
5	Reputation	The extent to which the information is trusted or highly regarded in terms of their source or content	25	Reputation of blog
			26	Reputation of bloggers
			27	Rewarding experiences
			28	Technorati rank
6	Info Representation	The way information is presented, maybe in different formats/media with customized displays	29	Exciting content
			30	Fresh perspective
			31	Multimedia
7	Navigation	The extent to which readers can move around the blog and retrieve information easily	32	Ease of ordering
			33	Easy to navigate
			34	Interactivity
8	Visual Design	Visual appearances that can attract readers	35	Attractive layout
			36	Clear layout of info
			37	Good use of colours
			38	Intuitive interface
9	Readability	Ability to comprehend the meaning of words or symbols	39	Easy to read info
			40	Legibility
			41	Readable font/text
10	Blog Accessibility	The extent to which the blog can be accessed faster and easier	42	Availability of info
			43	Blog responsiveness
			44	Ease of information access
11	Blog Technical Features	Features such as search tools, chat box, blogroll and comment field	45	Blogroll
			46	Chat box
			47	Comment field
			48	Search tool
			49	Trackback

4.2 Pilot Test

The statistics in Figure 6 summarize the responses in the pilot test by 40 persons (survey response = 100%) to 49 Likert-scale items covering the blog quality criteria. The mean of person ability estimates at +2.70 (SE .28) is the first indicator that blog readers find the pilot test comparatively easy, meaning they tend to accept all the proposed items. The standard deviation of 1.87 logits for person estimates indicates a greater spread in person variation than was observed in item-difficulty measures, which are even more restricted at .72. The person strata index of 6.54 (minimum person strata of 2) may provide information concerning the responsiveness of measures from this instrument and may be viewed as preliminary evidence for the external

aspect of construct validity. The mean-square fit and the *z* statistic are close to their expected values, +1 and 0, respectively, for items and persons, which demonstrates satisfactory fit to the model. The item reliability (Rasch equivalent to Cronbach's alpha) is .83 while person reliability is much higher at .98. Therefore, it can be inferred that (1) a line of inquiry has been developed in which some items are more difficult with respect to acceptance and some items are easier and (2) the consistency of these inferences can be expected. Similarly, it can be inferred that a line of inquiry has been developed in which some persons' levels of agreement are higher while others are lower and the consistency of these inferences can be expected. Reliability is the characteristic most commonly used in evaluating the generalizability aspect of construct validity. By substituting the person mean = +2.70 and item mean = 0 in Equation 1, we find the probability for acceptance of the 49 blog quality criteria by the blog readers is 93.7%, which exceeds the relative standard setting of Cronbach's alpha (70%). In this pilot test of the acceptability of blog quality criteria, designed as a screening device to identify the most-acceptable criteria to be used in blog quality assessment, this result provides crucial evidence to support the consequential aspect of construct validity.

Persons	40 INPUT	40 MEASURED		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	192.6	49.0	2.70	.28	1.00	-.1	.98	-.2
S.D.	29.7	.0	1.87	.06	.31	1.8	.31	1.8
REAL RMSE	.28	ADJ.SD	1.85	SEPARATION	6.54	Person RELIABILITY		.98
Items	49 INPUT	49 MEASURED		INFIT		OUTFIT		
MEAN	157.2	40.0	.00	.29	.99	-.1	.98	.0
S.D.	9.2	.0	.72	.01	.23	1.1	.24	1.0
REAL RMSE	.29	ADJ.SD	.66	SEPARATION	2.24	Item RELIABILITY		.83

Figure 6. Summary statistic

Figure 7 is a variable-map of pilot test analysis, showing the distribution of blog readers on the left and the distribution of item agreement on the right, according to person number and item label, respectively. The person and item distributions corroborate the results from the summary statistics.

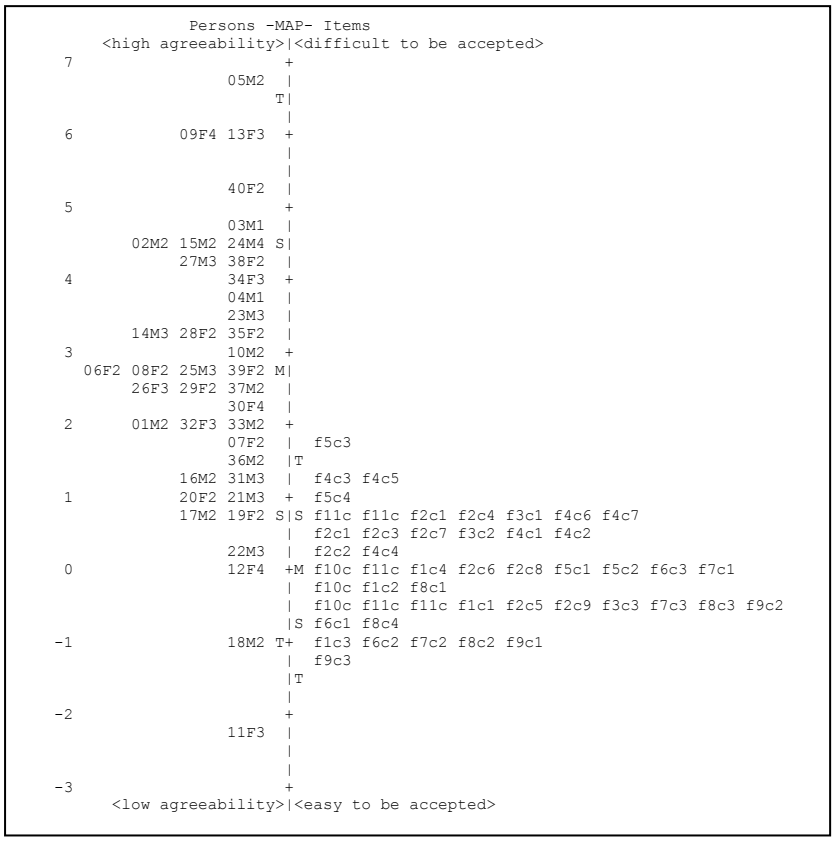

Figure 7. Variable maps

The outputs of the Rasch item and person estimates are listed in Figures 8 and 9, so the details of map locations can be verified conveniently. The easiest item in terms of acceptance is f9c3 (*Readable font*), located at the bottom of the item distribution at -1.30 logits (SE .30), while the most difficult item is f5c3 (*Rewarding experiences*), located at +1.75 logits (SE .27). The blog reader with the highest agreeability score is respondent 5, located at the top of the person distribution at +6.75 logits (SE .53), while the blog reader with the lowest agreeability score is respondent 11, located at -2.27 logits (SE .23). The fit statistics of the item output (see Figure 8) look very good, although we need to reconsider two overfit items, f10c1 (*Availability of info*) and f10c3 (*Ease of information access*). The Guttmann-like items do not cause any threat to measurement. Therefore, they are accepted for this analysis.

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
27	134	40	1.75	.27	1.31	1.4	1.31	1.3	.73	52.5	60.4	f5c3
20	140	40	1.31	.27	1.32	1.4	1.29	1.3	.59	60.0	60.7	f4c3
22	141	40	1.24	.27	1.36	1.5	1.35	1.5	.81	47.5	60.6	f4c5
28	143	40	1.09	.27	.93	-.3	.93	-.3	.72	62.5	60.7	f5c4
5	146	40	.87	.27	.82	-.8	.91	-.3	.74	62.5	60.7	f2c1
24	146	40	.87	.27	1.07	.4	1.11	.5	.69	57.5	60.7	f4c7
45	146	40	.87	.27	1.20	.9	1.22	1.0	.63	67.5	60.7	f11c1
46	146	40	.87	.27	1.19	.9	1.17	.8	.77	50.0	60.7	f11c2
23	147	40	.80	.27	1.19	.9	1.12	.6	.76	67.5	60.7	f4c6
8	149	40	.65	.27	1.28	1.2	1.27	1.2	.56	55.0	60.6	f2c4
15	149	40	.65	.27	1.16	.8	1.13	.6	.66	62.5	60.6	f3c1
16	150	40	.57	.27	1.31	1.4	1.28	1.2	.71	50.0	60.6	f3c2
18	150	40	.57	.27	1.35	1.5	1.39	1.6	.56	55.0	60.6	f4c1
7	151	40	.50	.27	.90	-.4	.86	-.6	.75	60.0	60.9	f2c3
19	151	40	.50	.27	.94	-.2	.91	-.3	.70	60.0	60.9	f4c2
11	152	40	.42	.27	1.31	1.4	1.31	1.3	.65	45.0	60.9	f2c7
14	152	40	.42	.27	1.00	.1	1.20	.9	.63	60.0	60.9	f2c10
21	154	40	.27	.28	.62	-1.9	.60	-1.9	.79	85.0	61.1	f4c4
6	155	40	.19	.28	1.09	.5	1.02	.2	.73	62.5	61.6	f2c2
12	156	40	.12	.28	.98	.0	1.18	.8	.68	60.0	61.7	f2c8
26	156	40	.12	.28	.99	.0	.95	-.1	.80	50.0	61.7	f5c2
49	156	40	.12	.28	.92	-.3	.87	-.5	.75	70.0	61.7	f11c5
10	157	40	.04	.28	.89	-.5	.83	-.7	.72	70.0	62.0	f2c6
25	158	40	-.04	.28	1.32	1.4	1.31	1.3	.61	52.5	62.4	f5c1
31	158	40	-.04	.28	1.20	.9	1.21	.9	.77	52.5	62.4	f6c3
32	158	40	-.04	.28	1.17	.8	1.13	.6	.68	52.5	62.4	f7c1
4	159	40	-.12	.28	1.11	.6	1.04	.3	.71	55.0	62.5	f1c4
43	159	40	-.12	.28	.73	-1.3	.72	-1.2	.73	72.5	62.5	f10c2
35	161	40	-.27	.28	.80	-.9	.75	-1.0	.78	62.5	62.6	f8c1
2	162	40	-.35	.28	.74	-1.2	1.11	.5	.83	75.0	62.8	f1c2
42	162	40	-.35	.28	.43	-3.3	.45	-2.7	.84	90.0	62.8	f10c1
9	163	40	-.43	.28	1.14	.7	1.06	.3	.70	50.0	63.1	f2c5
37	163	40	-.43	.28	.82	-.8	.80	-.8	.76	65.0	63.1	f8c3
44	163	40	-.43	.28	.53	-2.5	.52	-2.2	.85	80.0	63.1	f10c3
13	164	40	-.52	.29	1.07	.4	1.02	.2	.68	60.0	63.3	f2c9
17	164	40	-.52	.29	.72	-1.3	.67	-1.3	.85	72.5	63.3	f3c3
34	164	40	-.52	.29	.75	-1.2	1.12	.5	.76	72.5	63.3	f7c3
47	164	40	-.52	.29	.73	-1.3	.74	-1.0	.76	67.5	63.3	f11c3
1	165	40	-.60	.29	1.06	.3	.95	-.1	.75	65.0	63.5	f1c1
40	165	40	-.60	.29	1.25	1.1	1.21	.8	.71	57.5	63.5	f9c2
48	165	40	-.60	.29	.94	-.2	.92	-.2	.74	70.0	63.5	f11c4
29	168	40	-.85	.29	.72	-1.3	.65	-1.3	.80	72.5	64.4	f6c1
38	168	40	-.85	.29	.81	-.9	.78	-.7	.75	67.5	64.4	f8c4
3	169	40	-.94	.29	1.13	.7	1.01	.1	.70	67.5	65.1	f1c3
30	170	40	-1.02	.30	.67	-1.6	.63	-1.3	.78	72.5	65.4	f6c2
33	171	40	-1.11	.30	.92	-.3	.81	-.5	.76	70.0	65.9	f7c2
36	171	40	-1.11	.30	.91	-.3	.82	-.5	.74	75.0	65.9	f8c2
39	171	40	-1.11	.30	.88	-.5	.79	-.6	.77	70.0	65.9	f9c1
41	173	40	-1.30	.30	.80	-.9	.69	-.9	.77	72.5	66.7	f9c3

Figure 8. Item measures

Inspection of the Rasch fit statistics continues with the person measures displayed in Figure 9. Two persons, respondents 34 and 31, show somewhat erratic response patterns. After checking the section of the most misfitting response strings in Figure 10, the data are found to be noticeably unpredictable, but they do not degrade the measurement [28]. Hence, the two misfits are kept in the analysis. The item and person fit statistics can be used as direct evidence to support the substantive aspect of construct validity.

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
5	241	49	6.71	.53	1.04	.2	1.04	.3	.11	91.8	91.9	05M203C
9	238	49	6.06	.41	.95	.0	.81	-.4	.30	85.7	85.9	09F406A
13	237	49	5.90	.39	.84	-.5	.74	-.6	.42	85.7	83.9	13F306B
40	231	49	5.17	.32	1.18	.9	1.14	.6	.20	73.5	73.4	40F203A
3	227	49	4.80	.29	.67	-1.8	.63	-1.9	.60	77.6	67.3	03M101B
24	224	49	4.56	.28	1.11	.6	1.10	.6	.32	65.3	64.6	24M410C
2	223	49	4.48	.27	.96	-.1	.97	-.1	.32	63.3	63.8	02M203A
15	223	49	4.48	.27	1.11	.6	1.06	.4	.44	67.3	63.8	15M206D
38	221	49	4.33	.27	.88	-.6	.91	-.4	.22	55.1	62.1	38F201C
27	220	49	4.26	.27	1.09	.6	1.01	.1	.62	59.2	61.6	27M310D
34	215	49	3.92	.26	1.61	2.9	1.54	2.7	.34	55.1	59.6	34F304B
4	214	49	3.86	.25	.95	-.2	.92	-.4	.40	71.4	59.1	04M101B
23	210	49	3.61	.25	1.24	1.3	1.23	1.3	.39	42.9	57.7	23M306B
28	205	49	3.31	.24	1.36	1.8	1.33	1.7	.66	49.0	58.4	28F203A
35	205	49	3.31	.24	1.32	1.7	1.30	1.6	.69	42.9	58.4	35F204A
14	203	49	3.19	.24	.61	-2.4	.61	-2.4	.39	73.5	58.5	14M310B
10	199	49	2.96	.24	1.22	1.2	1.21	1.1	.38	53.1	58.9	10M203B
25	197	49	2.84	.24	.58	-2.6	.58	-2.6	.19	79.6	59.1	25M304C
39	197	49	2.84	.24	1.12	.7	1.12	.7	.57	46.9	59.1	39F203A
6	195	49	2.73	.24	.73	-1.5	.74	-1.4	.26	69.4	59.3	06F206B
8	195	49	2.73	.24	.34	-4.8	.34	-4.8	.45	85.7	59.3	08F206B
29	191	49	2.50	.24	.79	-1.1	.79	-1.1	.49	75.5	59.3	29F204B
37	190	49	2.45	.24	1.01	.1	1.02	.2	.58	55.1	59.3	37M204B
26	189	49	2.39	.24	.74	-1.4	.74	-1.4	.31	73.5	59.1	26F303A
30	188	49	2.33	.24	.98	.0	.98	.0	.67	57.1	58.9	30F404C
1	184	49	2.11	.24	1.02	.2	1.03	.2	.15	59.2	58.6	01M203A
33	182	49	2.00	.24	1.44	2.1	1.42	2.0	.39	49.0	58.4	33M203C
32	180	49	1.88	.24	.72	-1.5	.72	-1.6	.59	69.4	58.5	32F303B
7	178	49	1.77	.24	1.27	1.3	1.29	1.4	.14	53.1	58.5	07F206B
36	173	49	1.49	.24	.89	-.5	.90	-.4	.15	53.1	58.4	36M203C
31	170	49	1.31	.24	1.61	2.7	1.61	2.7	.82	51.0	58.6	31M310A
16	167	49	1.14	.24	.79	-1.0	.80	-1.0	.11	55.1	59.0	16M210A
20	165	49	1.02	.24	.85	-.7	.84	-.8	.16	61.2	59.2	20F201B
21	165	49	1.02	.24	1.29	1.4	1.32	1.5	.15	44.9	59.2	21M303C
19	161	49	.78	.24	1.43	1.9	1.46	2.0	.06	42.9	60.8	19F206A
17	160	49	.72	.24	1.31	1.4	1.32	1.5	.27	49.0	61.1	17M206A
22	154	49	.36	.25	.92	-.3	.92	-.3	.28	69.4	63.1	22M306A
12	147	49	-.07	.25	.17	-6.0	.16	-6.1	.00	98.0	64.8	12F406A
18	133	49	-.91	.24	.76	-1.1	.73	-1.3	.41	75.5	61.2	18M206B
11	108	49	-2.27	.23	.93	-.3	.92	-.4	.44	55.1	55.5	11F306A

Figure 9. Person measures

```
MOST MISFITTING RESPONSE STRINGS
Person          OUTMNSQ  |Item
                         |433 3244 43113 4 4 3321421 2111 111 2442 2222
                         |19638980174737922342150962614197865836545820 7
                      high------------------------------------------------
    31 31M310A     1.61 A|....55555.....................2222.........
    34 34F304B     1.54 B|............3.....3.....3..3.333...........5
```

Figure 10. Most misfitting response strings

In developing a high-quality measurement tool for blog assessments, the utility of rating scales should also be empirically investigated. Figure 11 represents the modelled category probability curve for item 1. Checks of category probability curves for other items show that they display the same curve. Observation of the expected succession of the curves' peaks verifies that the four thresholds are ordered and that there is a suitable distance between them. From this, it follows that the 5-point rating scale in the pilot test questionnaire yields highest-quality measures for the interest aspect of construct validity. The category probability curve is additional evidence for the substantive aspect of construct validity.

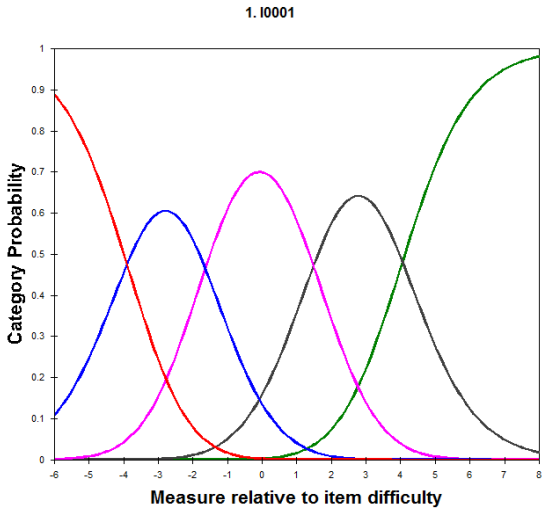

Figure 11. Category probability curve

Figure 12 shows a segment of principal contrast analysis of Rasch residual variance. The *variance explained by measures* is noticeably good (69.3%). The unidimensionality of the survey instrument is strongly confirmed by a more-likely-to-be-good *unexplained variance in the first contrast* (3.1%). This evidence of unidimensionality further supports the structural aspect of construct validity.

```
STANDARDIZED RESIDUAL VARIANCE SCREE PLOT
Table of STANDARDIZED RESIDUAL variance (in eigenvalue units)
                                          Empirical      Modelled
Total variance in observations     =        159.7 100.0%          100.0%
Variance explained by measures     =        110.7  69.3%           68.7%
Unexplained variance (total)       =         49.0  30.7% 100.0%    31.3%
Unexplained variance in 1st contrast =          5.0   3.1%  10.2%
Unexplained variance in 2nd contrast =          4.3   2.7%   8.8%
Unexplained variance in 3rd contrast =          3.2   2.0%   6.6%
Unexplained variance in 4th contrast =          3.2   2.0%   6.5%
Unexplained variance in 5th contrast =          2.8   1.8%   5.8%
```

Figure 12. Principal contrast analysis – Variance explained by measures should be >= 50% and unexplained variance in the first contrast should be <= 15% [29]

5 CONCLUSION AND FUTURE WORK

This article describes two tests designed to advance the development of a reliable instrument for assessment of blog quality. The content validity test investigated the acceptability of quality categories and criteria to expert reviewers, and the pilot test addressed the construct validity of the measurement instrument.

Rasch analyses provided empirical evidence of the criteria's construct validity in several aspects, including content, substantive, structural, generalizability, external and consequential aspects. The content validity test predicted expert reviewer agreement to definitions of 11 quality categories and 49 quality criteria assigned to those categories, after three criteria were removed for redundancy. The pilot test then confirmed that the criteria refined in the content validity test are accepted by blog readers. It is also confirmed that the Rasch measurement model is a powerful tool in evaluating construct validity.

The content validity and pilot tests are a crucial step toward development of a valid blog quality model. The tests ensure that our questionnaire provides meaningful measurements and that the content derived from our theoretical framework accords with blog readers' viewpoints with respect to blog quality.

This study does not establish the model for blog quality. Therefore, in future work, we plan to continue administering the revised questionnaire to further verify the acceptability of the blog quality criteria and thereby develop a significant blog quality model. The model will then be applied to create a blog quality assessment tool that can be used with high reliability in a wide variety of fields.

6 REFERENCES

1. Banks, M.A.: Blogging Heroes: Interviews with 30 of the World's Top Bloggers. Wiley Publishing Inc., Indianapolis, Indiana (2008).
2. Rowse, D. and Garrett, C.: ProBlogger: Secrets for Blogging Your Way to a Six-Figure Income. Wiley Publishing Inc., Indianapolis, Indiana (2008).
3. Tan, J.-E. and Ibrahim, Z.: Blogging and Democratization in Malaysia. A New Civil Society in the Making. SIRD, Petaling Jaya (2008).
4. Zhang, P., Dran, G.V., Blake, P., and Pipithsuksunt, V.: Important Design Features in Different Web Site Domains. e-Service Journal **1**(1), 77-91 (2001).
5. Katerattanakul, P. and Siau, K.: Information quality in internet commerce design. In: Piattini, M., Calero, C. and Genero, M. (eds.) Information and Database Quality, Kluwer Academic Publishers (2001).
6. Naumann, F. and Rolker, C.: Assessment methods for information quality criteria. In: Fifth International Conference on Information Quality (2000).
7. Graefe, G.: Incredible Information on the Internet: Biased information provision and a lack of credibility as a cause of insufficient information quality. In: Eighth International Conference on Information Quality, MIT, Cambridge, MA (2003).
8. Melkas, H.: Analyzing information quality in virtual service networks with qualitative interview data. In: Ninth International Conference on Information Quality (2004).
9. Eppler, M.: A generic framework for information quality in knowledge-intensive processes. In: Sixth International Conference on Information Quality. MIT, Cambridge, MA (2001).
10. Katerattanakul, P. and Siau, K.: Measuring Information Quality of Web Sites: Development of an Instrument. In: Proceeding of the 20th International Conference on Information System. GA, USA (1999).
11. Caro, A., Calero, C., Caballero, I., and Piattini, M.: A proposal for a set of attributes relevant for Web portal data quality. Software Quality J, **16**, 513-542 (2008).
12. Messick, S.: Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. American Psychologist, **50**(9), 741-749 (1995).
13. Kitchenham, B.A. and Pfleeger, S.L.: Personal Opinion Surveys. In: Shull, F., Singer, J., and Sjøberg, D.I.K. (eds.), Guide to Advanced Empirical Software Engineering. Springer-Verlag, London, 71-92 (2008).
14. Abdul Aziz, A., Mohamed, A., Arshad, N., Zakaria, S., Zaharim, A., Ahamd Ghulman, H., and Masodi, M.S.: Application of Rasch Model in validating the construct of measurement instrument. International Journal of Education and Information Technologies **2**(2), 105-112 (2008).
15. Rasch, G.: Weblogs models for some intelligence and Student test. The University of Chicago Press, Chicago (1980).
16. Masodi, M.S., Abdul Aziz, A., Mohamed, A., Arshad, N., Zakaria, S., and Ahamd Ghulman, H.: Development of Rasch-based Descriptive Scale in profiling Information Professionals' Competency. In: IEEE IT Simposium, Kuala Lumpur (2008).
17. Fisher, W.P., Jr.: The Rasch Debate: Validity and Revolution in Educational Measurement. In: Wilson, M. (ed.), Objective Measurement: Theory into Practice. Ablex, Norwood, NJ, 36-72 (1994).

18. Bond, T.G.: Validity and Assessment: A Rasch Measurement Perspective. Metodologia de las Ciencas del Comportamiento, **5**(2), 179-194 (2003).
19. Wolfe, E.W. and Smith, J., E. V.: Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part II - Validation Activities. Journal of Applied Measurement, **8**(2), 204-234 (2007).
20. Bond, T.V. and Fox, C.M.: Applying The Rasch Model: Fundamental Measurement in the Human Sciences. 2nd ed., Lawrence Erlbaum Associates, New Jersey (2007).
21. Sick, J.: Rasch Measurement in Language Education Part 3: The family of Rasch Models, Shiken. JALT Testing & Evaluation SIG Newsletter, **13**(1), 4-10 (2009).
22. Wright, B.D.: Rasch Model from Counting Right Answers: Raw Scores as Sufficient Statistics. Rasch Measurement Transactions, **3**(2), 62 (1989).
23. Wright, B.D. and Mok, M.M.C.: An overview of the family of Rasch measurement models. In: Everett, J., Smith, V., and Smith, R.M. (eds.), Introduction to Rasch Measurement: Theory, Models, and Applications. 979 (2004).
24. Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Test. Danish Institute for Educational Research, Copenhagen (1960).
25. Andrich, D.: A Rating Formulation for Ordered Response Categories. Psychometrika, **43**(4), 561-573 (1978).
26. Wright, B.D., Linacre, M., Gustafsson, J.-E., and Martin-Loff, P.: Reasonable Mean-square Fit Values. Rasch Measurement Transactions, **8**(3), 370 (1994).
27. Cronbach, L.J.: Coefficient Alpha and the Internal Consistency of Tests. Psychometrica, **16,** 297-334 (1951).
28. Linacre, J.M.: What do Infit and Outfit, MEan-square and Standardized mean? Rasch Measurement Transactions, **16**(2), 878 (2002).
29. Fisher, W.P., Jr.: Rating Scale Instrument Quality Criteria. Rasch Measurement Transactions, **21**(1), 1095 (2007).