JAMIA M1062 ONLINE DATA SUPPLEMENT APPENDIX ONE:

Effects of the Refined Quality Criteria on Reviews: A Study

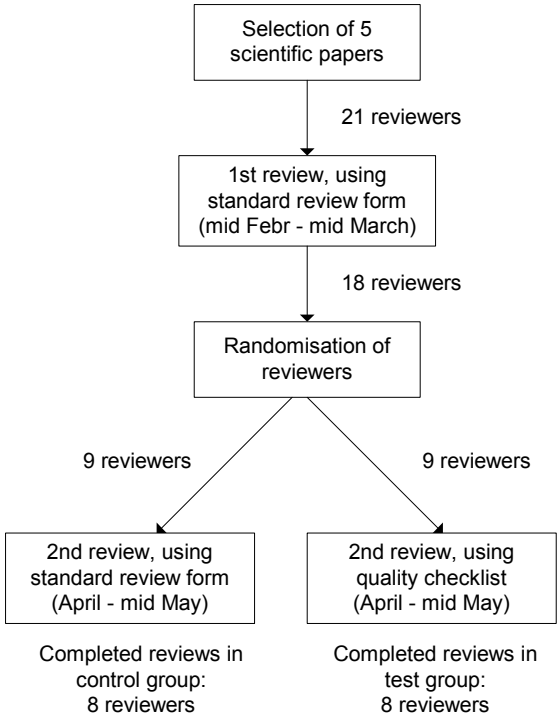
1. Study design

In order to evaluate the effects of the refined quality criteria on reviews, we conducted a randomized controlled trial. The aim was to answer the following three questions:
Q1 Does the variation between the reviewers change when the refined quality criteria are used?
Q2 Do the quantitative judgments of the reviewers change when the refined quality criteria are used?
Q3 Do the reviewers find the refined quality criteria useful to support reviews?
The study took place between February and May 2002. Twenty-one medical informatics researchers working in the area of information systems agreed to participate as reviewers in this study. Five papers were selected to be included in the study, taken as a sample from the about 60 candidate papers for the three information systems sections of the IMIA Yearbook 2002 [17].
The study was designed as a randomized controlled trial. The reviewers in the test group were asked to review each of the 5 papers twice: First with the usual 1-page evaluation form of the IMIA Yearbook with the five main quality criteria, and then again with the refined quality criteria. In order to be able to attribute any effect to the refined quality criteria, a control group of reviewers was defined which also evaluated each paper twice, but taking the 1-page evaluation form each time.
The distribution of reviewers to either the test group or control group was done randomly, stratified for their review experience. The washout time between the first and the second review was set to about 8 weeks. We considered this period sufficient to minimize the memory of details of the first review. In order to support this, all reviewed papers were re-collected after the 1st review, and the reviewers were asked not to keep a copy of the 1st review ratings. In order to guarantee that the reviewers of the test group really used the refined quality criteria during the second review, they were asked to check each item individually and note agreement or disagreement, before giving their overall rating for each category.
To answer Q1 on a change in the variation by the refined quality criteria, the t-test for paired samples was used to compare mean coefficients of variations between the first and second review in the test group.
To answer Q2 on a change in the mean ratings by the refined quality criteria, a four-way analysis of variance (ANOVA) with repeated measurements was used.
In both situations, to assure the applicability of the t-test and the analysis of variance, the Kolmogorov-Smirnov-Test with Lilliefors was used to check for normal distributions of the given ratings.
To answer Q3, the reviewers were asked to judge the usefulness of the refined quality criteria (only in the test group, after the 2nd review, using a Likert scale and open-ended questions).
All reviewers were also asked to document the time needed to complete the reviews, and to indicate the review experience (after the 1st review, using a Likert scale and asking for the years of review experience).

2. Execution of the study

The five selected papers covered various topics such as data mining, pharmacy system, clinical guidelines, computer-based reminders, and medication errors. Figure 1 shows the execution of the study. From the 21 reviewers who agreed to participate, 18 completed the first review. They were then randomized into the test group (9 researchers) and into the control group (9 researchers). 8 reviewers in each group completed the 2nd review. The reviewers who left the study gave as reason insufficient time to complete reviews.
From the eight reviewers in each group, five stated they had relatively little experience in reviewing papers (0 - 2 years of experience), and 3 stated that they were more experienced reviewers (5 – 15 years of experience). None of them had participated in the development of the refined quality criteria. The mean number of days between the first and second review of the same papers was 64 days ± 12 days, or about 2 months.

Figure 1: Flowchart of the study implementation for evaluating the effects of the quality checklist. Overall, 16 reviewers completed the study (8 in the test group, 8 in the control group).



3. Study results

i. Change of variation between reviewers

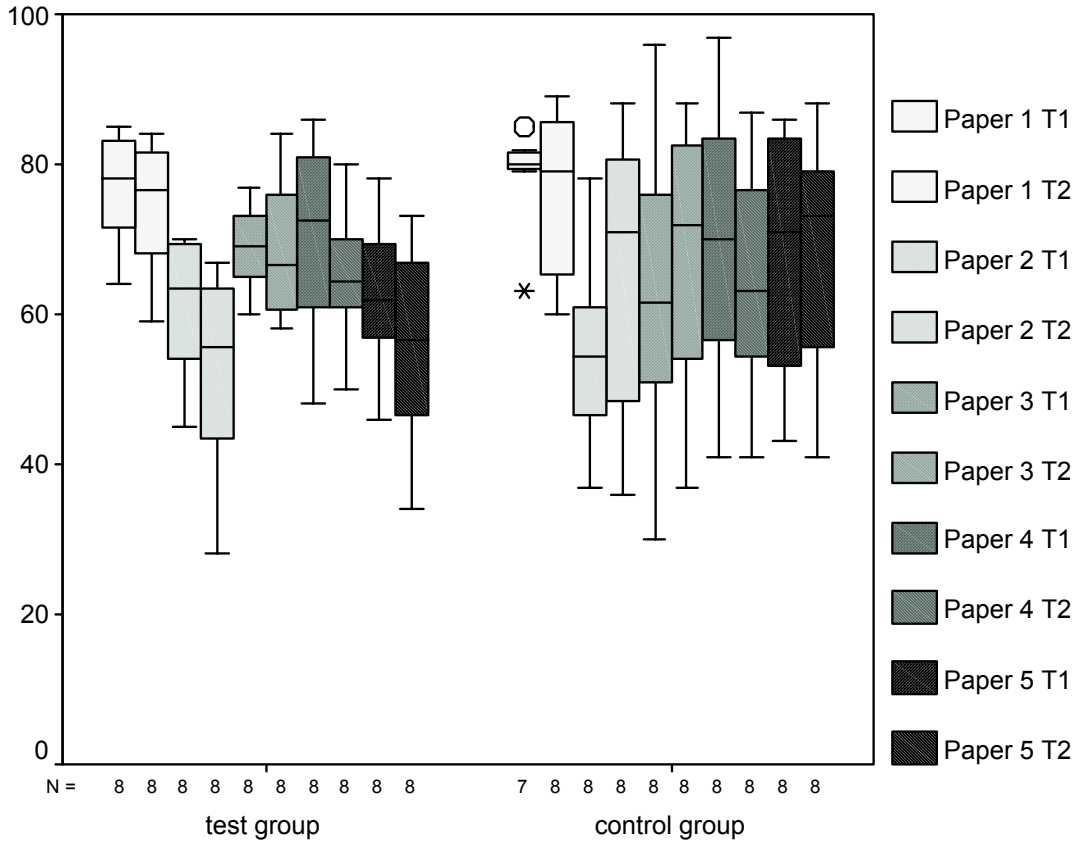
The variation index was chosen as an index for the variability of reviews. It was calculated for each question as standard deviation divided through the mean. In the control group, the variation index was 0.29 ± 0.12 at the 1st review, and 0.27 ± 0.09 in the 2nd review. In the test group, the variation index was 0.23 ± 0.08 and 0.24 ± 0.10, respectively.
Since it could be shown with the Kolmogorov-Smirnov-Test with Lilliefors correction that the variation indices were approximately normally distributed, the paired t-test could be used to check the hypothesis of a change in variation. The hypothesis could not be rejected. Thus, no reduction of variation between the reviewers in the test group could be confirmed by using the refined quality criteria.

ii. Change of mean ratings of papers

Figure 2 shows the mean ratings and distribution for each paper in the test group and in the control group. All papers have already been published and have thus been peer-reviewed, therefore it is not surprising that nearly all ratings are higher than 60 (from 100).

The Kolmogorov-Smirnow-Test with Lilliefors correction showed that the ratings are approximately normally distributed. Therefore, a 4-factor analysis of variance with repeated measurements could be conducted, using paper, time and question as within-subject factors, and group as between-subject factor. The results showed a significant interaction between paper, review time and group ($p = 0.027$), and between review time and group ($p = 0.034$). Post hoc analyses showed that the mean overall rating in the test group was significantly reduced from 68.0 ± 6.8 in the 1st review to 63.3 ± 4.1 in the 2nd review ($p = 0.001$), with the mean reduction being about 5 points (equivalent to 5%). In the control group, no significant changes could be found (66.2 ± 12.6 vs. 68.4 ± 14.0; $p = 0.485$).

Figure 2: Distribution of ratings for each of the 5 papers in the test group (8 reviewers) and in the control group (8 reviewers). T1 = 1st review, T2 = 2nd review, using box-whisker-plots. The maximum possible rating score is 100, indicating the highest quality, the lowest quality rating being 0.



Thus, the hypotheses of a change in mean ratings could not be rejected for the main factor effects group, review time and paper, but it could be rejected for the interaction of review time and group as well as for the interaction of review time, group and paper. The significant interaction between review time and group was constituted by lower (stricter) ratings in the test group and no change of the ratings in the control group.

iii. Usefulness of refined quality criteria in the opinion of reviewers

Table 2 shows the time needed to complete the review, as documented by the reviewers. In the control group the time needed to review the papers decreased by about 1/3 during the 2nd review, while in the test group, the review time increased by about 1/3. Both results are not surprising: The time needed to read and rate a paper is certainly lower when the paper has already been read and reviewed earlier, and higher when using the extended refined quality criteria instead of the 1-page standard criteria. Altogether, the mean review time was increased by about 50% when the refined quality criteria is thoroughly used.

Table 2: Mean time and standard deviation (in minutes) to terminate the review of the five papers during the 1st review and the 2nd review. The test group used the refined quality criteria during the 2nd review, otherwise the standard 1-page evaluation form was used.

	1st review	2nd review
Control group	32.6 ± 12.2	22.8 ± 10.2
Test group	25.4 ± 10.9	33.8 ± 17.7

Table 3 showed how the reviewers of the test group judged the usefulness of the refined quality criteria. In the free comments, the three more experienced reviewers remarked that the use of the refined quality criteria takes too much time (n=3), that some criteria cannot be applied to all papers (n=1), and that certain types of papers (such as innovative papers) cannot adequately be judged with this list (n=2). On the positive side, one experienced reviewer commented that subjective opinions can now be better justified.

Table 3: Usefulness of the refined quality criteria, as seen by the 8 reviewers of the test group, after the 2nd review, on a 5-point Likert scale (-- = absolutely not, - = rather not, -/+ = maybe, + = rather yes, ++ = absolutely yes).

		--	-	-/+	+	++
Less experienced reviewers (n=5)	Felt that list supported me in the review				2	3
	Will use list to support further reviews			1	4	
More experienced reviewers (n=3)	Felt that list supported me in the review		1	1	1	
	Will use list to support further reviews		2		1	

The five less experienced reviewers stated that the refined quality criteria support reviews when review experience is low (n=2), that it helps to become conscious of the criteria (n=2), to justify more negative ratings (n=1), and that it supports a structured review process (n=1).