

Rubriq White Paper:

The Science of the Rubriq Scorecard

summary of scorecard development status as of March, 2013

Introduction	2
Early phase development	3
Content Validation (CV)	4
Content Validation 1 (CV1): Managing Reviewers	4
Content Validation 2 (CV2): Managing Editors	7
Content Validation 3 (CV3): Academic Reviewers	8
Content Validation 4 (CV4): Editors-In-Chief of Research Journals	9
Novelty and Interest	11
Current Rubriq Scorecard (V6)	11
Weighting and Scoring	13
Further Scorecard Validation Studies: Scale Structure and Reliability	13
Descriptive Statistics	13
Correlations Among Scales	13
Internal Consistency Reliability	14
Independence of Reviewer Assessments	14
Conclusions From Scale Structure and Reliability	14
The Present and Moving Forward	14
References	16

Introduction

The Rubriq system of independent peer review uses a proprietary scorecard rating system designed to uniformly assess the quality of manuscripts intended for submission to peer reviewed research journals. The scorecard accommodates the ratings of three peer reviewers and offers quantitative scores for Quality of Research, Quality of Presentation, and Novelty and Interest in addition to a detailed commentary on quality.

The Rubriq scorecard is based on a standardized and validated set of metrics and is designed to offer uniformly consistent critiques of submitted papers. The initial focus will be on research reports in the areas of biological sciences and medicine. Three reviewers fill out individual scorecards. Data from the three scorecards comprise the final Rubriq scorecard report. The latter can be used by authors to improve their manuscript before journal submission and to find the highest impact journals that fit their papers. The Rubriq scorecard report and corresponding R-Score can also be used by journal Editors to identify and attract unpublished articles that match their specific criteria for acceptance.

The scorecard used to generate Rubriq consists of a three-part grid with separate scales for Quality of Research (QoR), Quality of Presentation (QoP), and Novelty and Interest (N&I).

Rubriq Scorecard System	
Quality of Research Scale (QoR)	Quality of Presentation Scale (QoP)
Subscales Hypothesis/Objective/Rationale Methods and Data Interpretation	Subscales Title page/abstract/introduction Results (text) Results (tables/figures) Discussion References Conclusions Writing (overall quality)
Novelty and Interest (N&I)	
Novelty Proof of principle New idea Proves an established idea	Interest Broad Moderate Limited

Each of the scales is comprised of subscales (Methods, Results, Discussion, etc.) for assessing quality. During the review process, reviewers indicate whether a manuscript has weaknesses with respect to each of the subscale items; if so, score deductions are automatically calculated. The aggregated scale scores are weighted (as described below) and then tallied to yield an overall final score for manuscript quality. In the following sections, we discuss the steps and processes used in developing the scorecard.

Early phase development

The following early phase development procedure was used to construct the Rubriq scorecard.

- A Scorecard Development Team (SDT) was formed to review the content of twenty reviewer guides used by biomedical research journals and to extract the most important, non-overlapping items that are essential to manuscript quality (data, hypothesis, methods, etc.). This list was further refined and collected under the headings of Quality of Research (QoR), Quality of Presentation (QoP), Novelty and Interest (N&I), and Significance.
- The early phase scorecard used ratings such as fair, good, etc. for each of the QoR and QoP attributes (V.1, example shown below). N&I and Significance focused on increasing levels of innovation and importance.

V.1																																																								
STANDARDIZED PEER REVIEW FORM – ORIGINAL RESEARCH ARTICLE																																																								
<p>Instructions: As a peer reviewer for this article, you play an invaluable role in assessing this research study based on your knowledge of the field. Your charge is threefold. First, to ensure the study is valid science and worthy of publication in any public repository. Second, to help categorize the article so it can be discovered by the target audience. Third, to provide some level of initial impact and novelty to the study to help the research community put this study in the context of previous research.</p> <p>Your review will be combined with other peer reviewers for this study to provide a final assessment and peer review score of this study.</p>																																																								
<p>PART 1: CLASSIFICATION ASSESSMENT</p> <p>1. Which fields would consider this study most relevant? DROPDOWN MENU: AREAS OF STUDY</p> <p>3. Who would find this interesting? DROPDOWN MENU: Clinicians, Policymakers, Academics, and Researchers</p>																																																								
<p>PART 2: GOOD SCIENCE ASSESSMENT</p> <p>The ratings in this section are designed to assess whether this research article is presenting a valid scientific study, and thus worthy of being included in the corpus of peer reviewed literature.</p> <table border="1"> <thead> <tr> <th>Minimum Publication Criteria</th> <th>Yes</th> <th>No</th> </tr> </thead> <tbody> <tr> <td>The study presents the results of primary scientific research.</td> <td></td> <td></td> </tr> <tr> <td>Results reported have not been published elsewhere.</td> <td></td> <td></td> </tr> <tr> <td>Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail.</td> <td></td> <td></td> </tr> <tr> <td>Conclusions are presented in an appropriate fashion and are supported by the data.</td> <td></td> <td></td> </tr> <tr> <td>The article is presented in an intelligible fashion and is written in standard English.</td> <td></td> <td></td> </tr> <tr> <td>The research meets all applicable standards for the ethics of experimentation and research integrity.</td> <td></td> <td></td> </tr> <tr> <td>The article adheres to appropriate reporting guidelines and community standards for data availability.</td> <td></td> <td></td> </tr> </tbody> </table>			Minimum Publication Criteria	Yes	No	The study presents the results of primary scientific research.			Results reported have not been published elsewhere.			Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail.			Conclusions are presented in an appropriate fashion and are supported by the data.			The article is presented in an intelligible fashion and is written in standard English.			The research meets all applicable standards for the ethics of experimentation and research integrity.			The article adheres to appropriate reporting guidelines and community standards for data availability.																																
Minimum Publication Criteria	Yes	No																																																						
The study presents the results of primary scientific research.																																																								
Results reported have not been published elsewhere.																																																								
Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail.																																																								
Conclusions are presented in an appropriate fashion and are supported by the data.																																																								
The article is presented in an intelligible fashion and is written in standard English.																																																								
The research meets all applicable standards for the ethics of experimentation and research integrity.																																																								
The article adheres to appropriate reporting guidelines and community standards for data availability.																																																								
<p>Quality of Research</p> <table border="1"> <thead> <tr> <th>Criteria</th> <th>Poor</th> <th>Weak</th> <th>Modest</th> <th>Strong</th> <th>Very Strong</th> </tr> </thead> <tbody> <tr> <td>Overall Quality of Data</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Hypothesis</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Methods & Experimental Design</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Conclusions</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> <p>Quality of Presentation</p> <table border="1"> <thead> <tr> <th>Criteria</th> <th>Poor</th> <th>Weak</th> <th>Modest</th> <th>Strong</th> <th>Very Strong</th> </tr> </thead> <tbody> <tr> <td>Writing Style</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>References</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Results</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> <p>Comments to Author on "Good Science" Criteria:</p> <div style="border: 1px solid black; height: 30px;"></div>			Criteria	Poor	Weak	Modest	Strong	Very Strong	Overall Quality of Data						Hypothesis						Methods & Experimental Design						Conclusions						Criteria	Poor	Weak	Modest	Strong	Very Strong	Writing Style						References						Results					
Criteria	Poor	Weak	Modest	Strong	Very Strong																																																			
Overall Quality of Data																																																								
Hypothesis																																																								
Methods & Experimental Design																																																								
Conclusions																																																								
Criteria	Poor	Weak	Modest	Strong	Very Strong																																																			
Writing Style																																																								
References																																																								
Results																																																								
<p>PART 3: IMPACT ASSESSMENT</p> <p>The ratings in this section will not determine whether the research article is published, but rather provide an initial indication of impact to the community upon publication. Post publication review will be available to the overall research community to continually assess impact of this research article over time.</p> <p>Novelty & Interest</p> <table border="1"> <thead> <tr> <th>Check</th> <th>Criteria</th> </tr> </thead> <tbody> <tr> <td></td> <td>Innovative finding of broad interest</td> </tr> <tr> <td></td> <td>Innovative contribution of limited appeal</td> </tr> <tr> <td></td> <td>Largely confirmatory and of general interest</td> </tr> <tr> <td></td> <td>Largely confirmatory and of specialized interest</td> </tr> <tr> <td></td> <td>Limited in scope and appeal</td> </tr> </tbody> </table> <p>Significance</p> <table border="1"> <thead> <tr> <th>Check</th> <th>Criteria</th> </tr> </thead> <tbody> <tr> <td></td> <td>Outstanding work of great significance</td> </tr> <tr> <td></td> <td>Good and useful advance in the field</td> </tr> <tr> <td></td> <td>Not a significant advance, but technically sound</td> </tr> <tr> <td></td> <td>Technically unsound</td> </tr> </tbody> </table> <p>Top 10</p> <table border="1"> <thead> <tr> <th></th> <th>Yes</th> <th>No</th> </tr> </thead> <tbody> <tr> <td>Would you rate this study in the top 10% of research you have read in the past year?</td> <td></td> <td></td> </tr> </tbody> </table> <p>Comments to Author on "Impact" Criteria:</p> <div style="border: 1px solid black; height: 30px;"></div>			Check	Criteria		Innovative finding of broad interest		Innovative contribution of limited appeal		Largely confirmatory and of general interest		Largely confirmatory and of specialized interest		Limited in scope and appeal	Check	Criteria		Outstanding work of great significance		Good and useful advance in the field		Not a significant advance, but technically sound		Technically unsound		Yes	No	Would you rate this study in the top 10% of research you have read in the past year?																												
Check	Criteria																																																							
	Innovative finding of broad interest																																																							
	Innovative contribution of limited appeal																																																							
	Largely confirmatory and of general interest																																																							
	Largely confirmatory and of specialized interest																																																							
	Limited in scope and appeal																																																							
Check	Criteria																																																							
	Outstanding work of great significance																																																							
	Good and useful advance in the field																																																							
	Not a significant advance, but technically sound																																																							
	Technically unsound																																																							
	Yes	No																																																						
Would you rate this study in the top 10% of research you have read in the past year?																																																								

To reduce variability in reviewer responses related to each attribute, more fully detailed descriptions were added to each category (V.2, QoR). The descriptors were intended to help a reviewer choose the most appropriate rating in each QoR and QoP subscale. Initial descriptors were based on our experience in assessing more than 1,000 consecutive manuscripts submitted for peer review to American Journal Experts (AJE) prior to journal submission. Also for V.2, the Significance scale was removed, and N&I was expanded to 9 levels ranging from poor to exceptional (V.2, N&I).

V.2					
Quality of Research					
Please evaluate the quality of the following:	Poor	Fair	Good	Very Good	Excellent
Hypothesis	Missing	Included but inconsistent with objectives or results	Poorly stated or missing from either the Abstract or Introduction; unreferenced in the Discussion	Well stated in Abstract and Introduction; addressed in the Discussion	Well stated, appropriately incorporated and captures the interest of the reader
Data	Does not support objectives and are incomplete	Supports objectives but are incomplete or unsubstantial	Supports objectives and are substantial but poorly presented	Supports objectives, moderately rigorous and well presented	Focused, rigorous, complete and convincing

V.2 Novelty & Interest								
Poor	Marginal	Fair	Satisfactory	Good	Very Good	Excellent	Outstanding	Exceptional
<ul style="list-style-type: none"> Limited in scope Largely confirmatory Of low interest to the field 			<ul style="list-style-type: none"> Advances the field Likely of wide interest Solid approach with useful data 			<ul style="list-style-type: none"> Address an important problem Built upon original insights or innovative techniques Will generate new research of interest to the field 		

Content Validation (CV)

Further development and refinement of the scorecard was carried out by a series of Content Validation studies. Content Validity refers to comprehensiveness, or to how adequately the sampled elements of an instrument are relevant to, and representative of, the targeted construct for a particular assessment purpose (Haynes, et al, 1995). For example: Do all of the items appear relevant to the concept being measured? Are the key aspects of manuscript quality covered under QoR, QoP and N&I? Content validity requires the input of subject matter experts (SME) to evaluate whether the questionnaire items adequately represent the construct the instrument is intended to address, which in the case of the scorecard is manuscript quality.

Content Validation 1 (CV1): Managing Reviewers

Content Validity was initially assessed by a panel of SMEs comprised of six Managing Reviewers from AJE (SME-MRs) who were not involved in the early phase development

of the scorecard. This test made use of a Likert Scale, which is a type of psychometric response scale that is widely used in survey research (Carifio and Perla, 2007; McDowell and Newell, 1996). When responding to a Likert questionnaire item, respondents specify their levels of agreement with a series of statements. Traditionally a five-point scale is used. An example of a section of the Likert Scale used in the CV1 test is shown here.

Likert Scale						
Content Validation Test 1		Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
		5	4	3	2	1
QoR and QoP						
1	The items in the QoR section appear appropriate for assessing the quality of the research.					
2	The items in the QoP section appear appropriate for assessing the quality of the manuscript.					

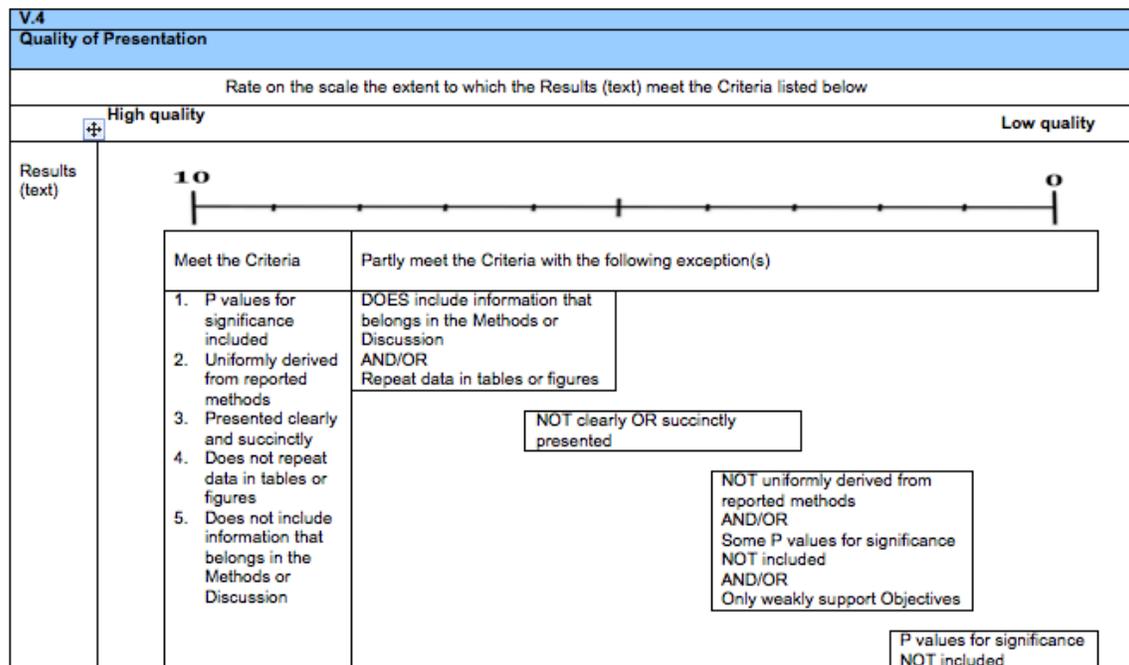
CV1 contained 16 statements related to the QoR, QoP, and N&I scales. The SME-MRs were asked to evaluate the adequacy of the descriptions of each subscale item. In all cases *Agree* or *Strongly Agree* indicated a favorable response because the questions were posed as positive statements, e.g., “The Scorecard is relevant to all users: peer reviewers, authors, journal editors”. *Undecided* is a neutral position. The panelists’ levels of agreement with the statements are coded 1-5 for scoring, which indicates increasing levels of agreement. Mean values were then calculated.

Likert Scale results for CV1 gave a mean summary score for QoR and QoP scale items equal to 4 out of a maximum value of 5, indicating there was general agreement that the descriptors represent increasing levels of quality and offer clear distinctions among choices. The scorecard received high marks for being consistent with its intended overall purpose and that it would be relevant for its intended audience of peer reviewers, authors, and journal editors.

The expert panel was encouraged to leave comments related to each response. The most frequently cited areas of needed improvement for descriptors concerned consistency of language, avoiding subjective statements, and the need for more uniform gradations from lower to higher quality. As a result, the descriptors were rewritten. An example from V.3 is shown below.

V.3					
Quality of Research					
	VERY GOOD	GOOD	ADEQUATE	FAIR	POOR
Hypothesis and Objective	<ul style="list-style-type: none"> Both included in abstract and introduction Are consistent with each other Supported by background information, or statement of rationale Consistent with experimental approach Consistent with, or addressed within, the discussion 	<ul style="list-style-type: none"> Both included in abstract and introduction Are consistent with each other Supported by background information or statement of rationale Consistent with experimental approach 	<ul style="list-style-type: none"> Both included in abstract and introduction Consistent with experimental approach AND EITHER <ul style="list-style-type: none"> Are NOT consistent with each other OR <ul style="list-style-type: none"> NOT Supported by background information or statement of rationale 	<ul style="list-style-type: none"> Missing from EITHER abstract and introduction OR <ul style="list-style-type: none"> Missing Hypothesis or Objective OR <ul style="list-style-type: none"> NOT supported by background information or statement of rationale OR <ul style="list-style-type: none"> NOT consistent with experimental approach 	<ul style="list-style-type: none"> Missing Hypothesis and Objective

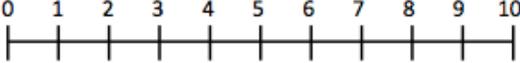
Subsequently, the SDT further clarified the language. The value categories (very good, good, etc.) were replaced by a Visual Analog Scale (VAS) anchored at 0 and 10 (V.4, below). A VAS can be used to measure characteristics that range across a continuum of values. The scale is conventionally a straight line, typically of 10 cm, anchored at each end with labels that indicate the range being considered (McDowell and Newell, 1996). Internet-based versions of the VAS have been shown to be valid (Reips and Funke, 2008). In our use of the VAS for the scorecard, the line is anchored at 0 (lowest quality) and 10 (highest quality). Subsection and section weightings were applied based on a reviewer survey (see below), and an overall score algorithm was established. Final weightings will be determined from data derived from field testing the scorecard.



To further reduce the subjective nature of the scoring, it was agreed to apply a percentage deduction system to all subscale items. The language was then rewritten as focused statements of weaknesses. Each statement was assigned a deduction corresponding to the severity of the weakness. As reviewers ticked off manuscript weaknesses, the score was reduced from a starting value of 10. Novelty and Interest ratings were changed from discreet values (1-9) to a 0-10 VAS anchored with descriptors at the high and low ends. Reviewers could then adjust the calculated score manually.

Content Validation 2 (CV2): Managing Editors

The scorecard underwent further Content Validation to assess the relevance and representativeness of the descriptors, and the appropriateness of the scoring system. Similar to the CV1 test, a 5-point Likert Scale was used for CV2; however, the subject panel was comprised of 13 AJE Managing Editors (SME-ME). The Likert Scale results indicated that the SME-ME *Agreed* or *Strongly Agreed* with statements concerning the adequacy of the scorecard for its intended purpose; however, they also offered suggestions for adding key statements to the list of weaknesses in the areas of data repetition, adequacy of the background information, and the consistency of data in text vs. tables. The N&I section was also further refined (a portion of V. 5.1 is shown below).

V.5.1			
Novelty and Interest			
Irrespective of the data presented and the presentation of the paper, how interesting is the research question or objective of this paper?			
0 1 2 3 4 5 6 7 8 9 10 			
Low		High	
<ul style="list-style-type: none"> Limited in scope Largely confirmatory Of low interest to the field 		<ul style="list-style-type: none"> Effectively advances understanding of an important problem Built upon original insights or innovative techniques Will generate new research of interest to the field 	
Check the items that apply to each subsection and a suggested starting score (0-10) will be generated. You will be able to adjust this score along the scale as you see fit.			
Quality of Research		% deduction	select
Hypothesis, Objective, Rationale	Manuscript errors listed below do not apply	0%	
	Not consistent with discussion	30%	
	Missing from abstract	35%	
	Missing from introduction	35%	
	Not supported by background/introduction	50%	
	Not consistent with experimental approach	70%	
Methods and Data	Manuscript errors listed below do not apply	0%	
	Missing some references	20%	
	Includes unessential data	20%	
	Insufficient sample size	25%	
	Missing important details for reproducibility	30%	
	Missing some experimental controls	55%	
	Inappropriate statistical analyses	55%	
	Approach/data not consistent with objectives	70%	

Content Validation 3 (CV3): Academic Reviewers

A third Content Validity assessment was conducted in which the subject matter expert panel was comprised of 21 independent, university-based scientists who serve as peer reviewers for AJE manuscript submissions. The test consisted included the Likert Scale questionnaire used previously and a Lawshe Scale. Lawshe (1975) created a Content Validity Ratio (CVR) that is used to gauge the content validity of items on an empirical measuring system. In this method, a panel is asked to indicate whether or not a measurement item in a set is *Essential* to the operationalization of a construct. In the present case, the constructs are represented by QoR, QoP, and N&I; the measurement items are the subscale weaknesses. “*Essential*” items are those that best represent the goal of the scorecard, which is to measure the quality of manuscripts. Other choices in the Lawshe Scale are “*Useful, but not essential*” and “*Not necessary*”.

The Lawshe CVR = $(2n_e / N) - 1$, where n_e is the number of SMEs who believe the item is essential and N is the total number of SMEs. The CVR yields values between -1.00 and +1.00. A CVR of 0.00 indicates that 50% of the SMEs in a panel size of N believe the attribute is essential and has some degree of Content Validity. The higher the percentage, the greater the validity. There are established minimum CVRs for varying panel sizes based on a one tailed test at the $\alpha=0.05$ significance level. For a panel of 21 reviewers, the critical value is >0.359 (Wilson, et al, 2012). A portion of the Lawshe Scale used for CV3 is shown here.

Lawshe Scale				
Quality of Presentation		Essential	Useful, but not essential	Not necessary
Results (text)	Excessive repetition of data (e.g., from tables or figures)			
	Not clearly/succinctly presented			
	Some results utilize methods not explained in the methods section			
	Missing significance indicators (e.g., values-values)			
	Not focused on objectives			
Results (tables and figures)	Problems with ordering, numbering, titles, or labels			
	Present unnecessary, or superfluous information			
	Captions/footnotes/legends are incomplete or missing			
	Missing important data			

Nearly 90% of the 42 subscale items were endorsed by more than 50% of panelists as representing essential items for assessing the quality of manuscripts; four items were viewed as useful but not essential, and these items were modified or eliminated. Only one descriptor was considered not necessary (“Self-selection bias” for references) and was removed. Panelists were also asked for their qualitative assessments of individual item weaknesses. Based on a wealth of helpful responses, the development team

further refined the language. The SDT also moved the Conclusions subscale to the QoP scale and added a new subscale to QoP, viz., “Title page, abstract, introduction”. That subscale is shown here as seen in scorecard version 5.2:

V.5.2	
Quality of Presentation	Starting value=10
Title page is incomplete	<i>Each weakness selected by a reviewer is associated with a percentage score reduction</i>
Hypothesis/objective is missing from abstract or introduction	
Background/rationale is missing from abstract or introduction	
Methods are missing from abstract	
Results are missing from abstract	
Conclusions are missing from abstract	

Content Validation 4 (CV4): Editors-In-Chief of Research Journals

Nineteen Editors-In-Chief (EIC) of academic science journals participated in CV4. The Likert Scale for the Content Validity exercise was comprised of 21 items drawn from papers written by journal editors or researchers concerned with the peer review process. We analyzed the following bibliography to identify the key factors that journal editors want to see in high quality reviews by their reviewers. Those requirements were used to construct the Likert Scale for CV4.

Citations Consulted For Preparation of the Likert Scale for Content Validation 4	
Evaluating Peer Reviews: Pilot Testing of a Grading Instrument JAMA. 1994;272(2):98. http://jama.jamanetwork.com/article.aspx?articleid=376106	Validation of an index of the quality of review articles J Clin Epi Volume 44, Issue 11, 1991, Pages 1271–1278. http://www.sciencedirect.com/science/article/pii/089543569190160B
Measuring the Quality of Editorial Peer Review JAMA. 2002;287(21):2786-2790. doi:10.1001/jama.287.21.2786.	Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts Journal of Clinical Epidemiology J Clin Epi Volume 52, Issue 7 , Pages 625-629, July 1999. http://www.jclinepi.com/article/S0895-4356(99)00047-5/abstract
What Makes a Good Reviewer and a Good Review for a General Medical Journal? JAMA. 1998;280(3):231-233. doi:10-1001/pubs.JAMA-ISSN-0098-7484-280-3-jpv71033.	
Quality Assessment of Reviewers' Reports Using a Simple Instrument	Implementation of a Journal Peer Reviewer Stratification System Based on Quality and

Obs Gyn 2006 - Volume 108 - Issue 4 - pp 979-985. doi: 10.1097/01.AOG.0000231675.74957.48.	Reliability Ann Emerg Med Volume 57, Issue 2 , Pages 149-152.e4, February 2011. http://www.annemergmed.com/article/S0196-0644(10)01355-7/abstract
---	---

The outcomes of CV4 demonstrated general approval and support for the scorecard. The percent responses from the EICs indicated more than 70% endorsement of Likert Scale items listed as Strongly Agree + Agree, whereas less than 20% of Neutral items were selected.

This endorsement was emphasized by written comments offered by participating EIC:

Editors-In-Chief Comments on the Scorecard
“This is clearly a positive innovation. Its strength is in encouraging a set of uniform standards, which would improve the quality of weak reviews.”
“Useful addition to the arena of peer review.”
“I think this is a great step forward and would be extremely useful.”
“It is important to remind reviewers of all these points and give them a scoring system at hand. It will certainly improve the reviewing process.”
“Authors would benefit undoubtedly from the feedback, and from their point of view this would be an invaluable exercise.”
“The scorecard handles well the common elements of a paper but the comments sections allows for addressing the unique elements, so I think this approach will work just fine.”

The EIC suggested the inclusion of an overview commentary section at the top of the scorecard where it will be seen first by authors.

Novelty and Interest

The N&I scale also went through a series of modifications to improve clarity and preciseness of definitions and of scaling. As with the QoR and QoP scales, our goal with N&I is to reduce, as much as possible, subjective scoring and reviewer bias. Initially we proposed using an ordinal scale:

Novelty & Interest	
Impact	Quality
high <ul style="list-style-type: none"> Addresses an important problem Employs innovative techniques Challenges existing paradigms 	Exceptional
	Outstanding
	Excellent
medium <ul style="list-style-type: none"> Good and useful advance in the field Of sufficient interest 	Very Good
	Good
	Satisfactory
low <ul style="list-style-type: none"> Limited in scope Largely confirmatory Not a significant advance 	Fair
	Marginal
	Poor

However, this system was considered to be insufficiently quantitative. The qualitative attributes were replaced by a ten-point VAS anchored at the ends. Ultimately we elected to change the format to make the N&I assessment more specifically reflective of quality manuscripts. Novelty was separated from Interest, and the quality attributes were more specifically defined and given rankings, as shown in the current scorecard below (V6). Novelty and Interest scales are each assigned a maximum score of 10, and the scores are averaged to yield an overall N&I score.

Current Rubriq Scorecard (V6)

Based on the above considerations, the revised Rubriq scorecard is shown on the following page:

Novelty and Interest

Novelty

- New technique, method, or approach (proof of principle)
- New question, theory, or hypothesis (totally new idea)
- New result, discovery, or perspective/synthesis (proves an established idea)

Interest

- Of broad interest to researchers in this field and other fields
- Of broad interest within the field
- Of moderate interest within the field
- Of interest to a small group within the field
- Of limited interest

Quality of Research

Hypothesis, Objective, Rationale

- Rationale unclear
- Objective/hypothesis is not supported by background
- Objective/Hypothesis is not stated

Methods and Data

- Missing essential references
- Design/Techniques not up to date
- Missing important details for reproducibility
- Missing some experimental controls
- Inappropriate statistical analyses
- Missing an important experiment
- Approach/data not consistent with objectives/hypothesis

Interpretation

- Does not adhere closely to the data
- Biased or overstated interpretation
- Leads to inaccurate conclusions
- Not supported by the data

Quality of Presentation

Title, Abstract, and Introduction

- Title is inappropriate
- Background/rationale is missing from abstract or introduction
- Methods are missing from abstract
- Conclusions are missing from abstract

- Hypothesis/objective is missing from abstract or introduction
- Results are missing from abstract

Results (text)

- Excessive repetition of data (e.g., from tables or figures)
- Poorly organized or not succinctly presented
- Missing significance indicators (e.g., p-values)
- Not focused on objectives

Results (Tables and Figures)

- Problems with ordering, numbering, titles, or labels
- Data presentation is inappropriate for the experiments performed
- Images are of poor quality
- Figures are too complex, confusing, or unclear
- Captions, footnotes, or legends are incomplete or missing
- Missing important data

Discussion

- Missing concise and accurate summary of results
- Missing discussion of potential limitations
- Biased commentary
- Insufficient comparison to relevant literature/previous results
- Not consistent with objective/hypothesis

Conclusions

- Missing take-away statements
- Vague, overstated or understated applicability
- Not consistent with discussion or objectives
- Not supported by the data

References

- Problems with ordering or numbering
- Not focused on current/latest literature
- Do not provide informative context for Introduction/Discussion
- Do not support objectives
- Missing key references

Writing (overall quality)

- Incorrect level of depth in some or all sections (too brief or too lengthy)
- Poor/unscientific word choice or grammar
- Writing lacks clarity, focus, or organization

Weighting and Scoring

Weighted scores are determined for the individual subscales in QoR and QoP and converted to a 0 to 10 scale. An algorithm is then used to calculate the final score. Weightings are based on feedback from 87 independent academic researchers from multiple research disciplines. The average weighted summary score for the QoR and QoP scales is reported to authors along with an average N&I score. The quality scores of individual subscales are also shown in the authors' Rubriq report to pinpoint those areas of their manuscripts that need improvement. All scorecard items are matched to concise reviewer commentaries.

Further Scorecard Validation Studies: Scale Structure and Reliability

The scorecard was tested using actual reports submitted for pre-publication peer review. Three reviewers independently provided scorecard ratings and comments for each manuscript. Data for analysis were available for 47 manuscripts with three reviews per manuscript. Items from all subscales were included in the studies. The three reviewer-specific scores per manuscript were averaged and an overall score for QoR, QoP, and N&I was obtained. Detailed results will be presented elsewhere.

Analyses were conducted on the following:

- Descriptive statistics
- Internal consistency reliability of the scale scores
- Inter-rater reliability of the scale scores

Descriptive Statistics

The means, standard deviations, and ranges of the item and scale scores suggested that raters used the full spectrum of the 0 to 10 scales. The ranges of assigned scores were fairly wide and the means tended to fall close to the center of the scales. There did not appear to be any trends indicating that raters were consistently assigning low or high scores. Further, there were very few instances in which all 3 raters assigned the lowest or highest possible score to an item. A visual inspection of stem-and-leaf plots (not shown) confirmed the relatively normal distributions of the item and scale scores. These findings suggest that the items and scales can accommodate variation in manuscript quality.

Correlations Among Scales

A correlation matrix revealed statistically significant correlations ($p < 0.05$) among scale items. The following trends were noted:

- For the N&I scale, Novelty is strongly correlated with Interest and is less correlated with items from QoR or QoP

- Items within QoR are moderately correlated with each other
- Within the QoP scale, most correlations among the items were moderately high
- The inter-relatedness of the QoR and QoP was scales noted by the overall significant correlation between scales

Internal Consistency Reliability

We examined internal consistency, which refers to the consistency, or degree of correlation, among items in a subscale and which is a pre-requisite for validity. The estimated internal consistency reliabilities (Cronbach's alpha) were high (>0.75) for each of the three scales (QoR, QoP, and N&I).

Independence of Reviewer Assessments

When developing a measure that multiple raters will use to assign scores, the typical goal is to develop an instrument that is reliable across raters, i.e., provides the same or similar results among raters. In the case of reviewing manuscripts, however, this may not be desirable since multiple opinions are sought to ensure the most comprehensive evaluation. We examined whether or not the scorecard accommodates diverse opinions. The inter-rater reliability of the item and scale scores was assessed using a measure called the intraclass correlation coefficient (ICC). Large ICCs indicate greater agreement between raters, where 1.0 represents perfect agreement or reliability. In the case of the scorecard the average ICC for all item and scale scores was less than 0.25 indicating a diversity of scores assigned by 3 raters per manuscript. This may in part be due to a scorecard innovation that allows Reviewers to adjust the suggested calculated scores during the review process.

Conclusions From Scale Structure and Reliability

These exploratory analyses of Rubriq scorecard ratings assigned to 47 manuscripts provide an assessment of the QoR, QoP, and N&I scale scores and their individual items:

- The item and scale scores have relatively normal distributions and adequately capture variations in manuscript quality
- Raters are using the full spectrum of the 0 to 10 scales
- Ceiling and floor effects were not observed
- The scale structure accommodates a diversity of opinions by 3 raters for each manuscript
- The scales have high internal consistency reliability

The Present and Moving Forward

The scorecard development and validation studies were based on the standards set by journals for the quality of peer reviews, feedback from academic researchers, and consensus views of panels of subject matter experts. The items listed for QoR, QoP, and N&I have undergone an extensive quality review process to establish their relevance, accuracy, and uniqueness. At each step of the development process, the scorecard has been subjected to content validation by experts using established methods of survey

research. Internal consistency is a pre-requisite for validity, and our findings suggest that the QoR, QoP, and N&I scales meet this requirement.

Validation is an ongoing process as we further refine the descriptors, weighting, and the scoring systems. This process will incorporate the results of field tests using the current and subsequent versions of the scorecard and further validity assessments.

References

- Carifio, J. and Perla, R.J. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths And Urban Legends About Likert Scales And Likert Response Formats And Their Antidotes. *Journal of Social Sciences* 3, 106-116.
- Flynn, D., van Schaik, P. and van Wersch, A. (2004). A Comparison Of Multi-Item Likert And Visual Analogue Scales For The Assessment Of Transactionally Defined Coping Function. *European Journal of Psychological Assessment* 20, 49-58.
- Haynes, S. N., Richard, D. C. S., and Kubany, E. (1995). Content Validity In Psychological Assessment: A Functional Approach To Concepts And Methods. *Psychological Assessment* 7, 238-247.
- Lawshe, C. H. (1975). A Quantitative Approach To Content Validity. *Personnel Psychology* 28, 563–575.
- Likert, R. (1932). A Technique For The Measurement Of Attitudes. *Archives of Psychology* 22, 55.
- McDowell, I., and Newell, C. (1996). *Measuring Health*. Oxford: Oxford University Press.
- Reips, U.-D. and Frederik, F. (2008) Interval-level Measurement With Visual Analogue Scales In Internet-Based Research: VAS Generator. *Behavior Research Methods* 40, 699-704.
- Wilson, R.F., Pan, W., and Schumsky, D.A. (2012). Recalculation Of The Critical Values For Lawshe’s Content Validity Ratio. *Measurement and Evaluation in Counseling and Development*. DOI: 10.1177/0748175612440286.