Argumentation Quality Assessment: Theory vs. Practice

Henning Wachsmuth * Nona Naderi ** Ivan Habernal *** Yufang Hou ****
Graeme Hirst ** Iryna Gurevych *** Benno Stein *

* Bauhaus-Universität Weimar, Weimar, Germany, www.webis.de
** University of Toronto, Toronto, Canada, www.cs.toronto.edu/compling
*** Technische Universität Darmstadt, Darmstadt, Germany, www.ukp.tu-darmstadt.de
**** IBM Research, Dublin, Ireland, ie.ibm.com

Abstract

Argumentation quality is viewed differently in argumentation theory and in practical assessment approaches. This paper studies to what extent the views match empirically. We find that most observations on quality phrased spontaneously are in fact adequately represented by theory. Even more, relative comparisons of arguments in practice correlate with absolute quality ratings based on theory. Our results clarify how the two views can learn from each other.

1 Introduction

The assessment of argumentation quality is critical for any application built upon argument mining, such as debating technologies (Rinott et al., 2015). However, research still disagrees on whether quality should be assessed from a theoretical or from a practical viewpoint (Allwood, 2016).

Theory states, among other things, that a cogent argument has acceptable premises that are relevant to its conclusion and sufficient to draw the conclusion (Johnson and Blair, 2006). Practitioners object that such quality dimensions are hard to assess for real-life arguments (Habernal and Gurevych, 2016b). Moreover, the normative nature of theory suggests absolute quality ratings, but in practice it seems much easier to state which argument is more convincing—a relative assessment. Consider two debate-portal arguments for “advancing the common good is better than personal pursuit”, taken from the corpora analyzed later in this paper:

**Argument A** “While striving to make advancements for the common good you can change the world forever. Allot of people have succeeded in doing so. Our founding fathers, Thomas Edison, George Washington, Martin Luther King jr, and many more. These people made huge advances for the common good and they are honored for it.”

**Argument B** “I think the common good is a better endeavor, because it’s better to give then to receive. It’s better to give other people you’re hand out in help then you holding your own hand.”

In the study of Habernal and Gurevych (2016b), annotators assessed Argument A as more convincing than B. When giving reasons for their assessment, though, they saw A as more credible and well thought through; that does not seem to be too far from the theoretical notion of cogency.

This paper gives empirical answers to the question of how different the theoretical and practical views of argumentation quality actually are. Section 2 briefly reviews existing theories and practical approaches. Section 3 then empirically analyzes correlations in two recent argument corpora, one annotated for 15 well-defined quality dimensions taken from theory (Wachsmuth et al., 2017a) and one with 17 reasons for quality differences phrased spontaneously in practice (Habernal and Gurevych, 2016a). In a crowdsourcing study, we test whether lay annotators achieve agreement on the theoretical quality dimensions (Section 4).

We find that assessments of overall argumentation quality largely match in theory and practice. Nearly all phrased reasons are adequately represented in theory. However, some theoretical quality dimensions seem hard to separate in practice. Most importantly, we provide evidence that the observed relative quality differences are reflected in absolute quality ratings. Still, our study underpins the fact that the theory-based argumentation quality assessment remains complex. Our results do not generally answer the question of what view of argumentation quality is preferable, but they clarify where theory can learn from practice and vice versa. In particular, practical approaches indicate what to focus on to simplify theory, whereas theory seems beneficial to guide quality assessment in practice.
A fallacy is a kind of error that undermines reason-
Argumentation theory discusses logical, rhetorical,
This section outlines major theories and practical
Table 1: The 15 theory-based quality dimensions
cally into cogency, effectiveness, reasonableness,
Effectiveness
Credibility
Emotional appeal
Clarity
Appropriateness
Arrangement
Reasonableness
Global acceptability
Global relevance
Global sufficiency
Overall quality

Table 1: The 15 theory-based quality dimensions rated in the corpus of Wachsmuth et al. (2017a).

2 Theory versus Practice

2.1 Theoretical Views of Quality Assessment

Argumentation theory discusses logical, rhetorical, and dialectical quality. As few real-life arguments are logically sound, requiring true premises that deductively entail a conclusion, cogency (as defined in Section 1) is largely seen as the main logical quality (Johnson and Blair, 2006; Damer, 2009; Govier, 2010). Toulmin (1958) models the general structure of logical arguments, and Walton et al. (2008) analyze schemes of fallacies and strong arguments. A fallacy is a kind of error that undermines reasoning (Tindale, 2007). Strength may mean cogency but also rhetorical effectiveness (Perelman and Olbrechts-Tyteca, 1969). Rhetoric has been studied since Aristotle (2007) who developed the notion of the means of persuasion (logos, ethos, pathos) and their linguistic delivery in terms of arrangement and style. Dialectical quality dimensions resemble those of cogency, but arguments are judged specifically by their reasonableness for achieving agreement (van Eemeren and Grootendorst, 2004).

Wachsmuth et al. (2017a) point out that dialectical builds on rhetorical, and rhetorical builds on logical quality. They derive a unifying taxonomy from the major theories, decomposing quality hierarchically into cogency, effectiveness, reasonableness, and subdimensions. Table 1 lists all 15 dimensions covered. In Section 3, we use their absolute quality ratings from 1 (low) to 3 (high) annotated by three experts for each dimension of 304 arguments taken from the UKPConvArg1 corpus detailed below.

2.2 Practical Views of Quality Assessment

There is an application area where absolute quality ratings of argumentative text are common practice: essay scoring (Beigman Klebanov et al., 2016). Persing and Ng (2015) annotated the argumentative strength of essays composing multiple arguments with notable agreement. For single arguments, however, all existing approaches that we are aware of assess quality in relative terms, e.g., Cabrio and Villata (2012) find accepted arguments based on attack relations, Wei et al. (2016) rank arguments by their persuasiveness, and Wachsmuth et al. (2017b) rank them by their relevance. Boudry et al. (2015) argue that normative concepts such as fallacies rarely apply to real-life arguments and that they are too sophisticated for operationalization.

Based on the idea that relative assessment is easier, Habernal and Gurevych (2016b) crowdsourced the UKPConvArg1 corpus. Argument pairs (A, B) from a debate portal were classified as to which argument is more convincing. Without giving any guidelines, the authors also asked for reasons as to why A is more convincing than B. In a follow-up study (Habernal and Gurevych, 2016a), these reasons were used to derive a hierarchical annotation scheme. 9111 argument pairs were then labeled with one or more of the 17 reason labels in Table 2.
We now report on experiments that we performed where both have been annotated by Wachsmuth. As both corpora described in Section 2 are based on the UKPConvArg1 corpus and thus share many

correlations. Analyzing the single values, we find

difference—this is underlined by the generally high

correlations. The high \( \tau \)'s of 8-5 (more credible) for local acceptability (.73) and of 9-4 (well thought through) for cogency (.75) confirm the match assumed in Section 1. Also, the values of 5-3 (unclear) for clarity (.91) and of 7-2 (non-sense) for reasonableness (.94) as well as the weaker correlation of 8-4 (objective) for emotional appeal (.35) makes sense.

Only the comparably low \( \tau \) of 6-1 (no credible evidence) for local acceptability (.49) and credibility (.52) seem really unexpected. Besides, the descriptions of 6-2 and 6-3 sound like local but cor-

Table 3 presents all \( \tau \)-values. The phrasing of a reason can be assumed to indicate a clear quality difference—this is underlined by the generally high correlations. Analyzing the single values, we find much evidence for Hypothesis 1: Most notably, label 5-1 perfectly correlates with global acceptability, fitting the intuition that abuse is not acceptable.

The high \( \tau \)'s of 8-5 (more credible) for local acceptability (.73) and of 9-4 (well thought through) for cogency (.75) confirm the match assumed in Section 1. Also, the values of 5-3 (unclear) for clarity (.91) and of 7-2 (non-sense) for reasonableness (.94) as well as the weaker correlation of 8-4 (objective) for emotional appeal (.35) makes sense.

Only the comparably low \( \tau \) of 6-1 (no credible evidence) for local acceptability (.49) and credibility (.52) seem really unexpected. Besides, the descriptions of 6-2 and 6-3 sound like local but cor-

Table 3: Kendall’s \( \tau \) rank correlation of each of the 15 quality dimensions of all argument pairs annotated by Wachsmuth et al. (2017a) given for each of the 17+1 reason labels of Habernal and Gurevych (2016a). Bold/gray: Highest/lowest value in each column. Bottom row: The number of labels for each dimension.

by crowd workers (UKPConvArg2). These pairs represent the practical view in our experiments.

3 Matching Theory and Practice

We now report on experiments that we performed to examine to what extent the theory and practice of argumentation quality assessment match.1

3.1 Corpus-based Comparison of the Views

Several dimensions and reasons in Tables 1 and 2 seem to refer to the same or opposite property, e.g., clarity and 5-3 (unclear). This raises the question of how absolute ratings of arguments based on theory relate to relative comparisons of argument pairs in practice. We informally state three hypotheses:

Hypothesis 1 The reasons for quality differences in practice are adequately represented in theory.

Hypothesis 2 The perception of overall argumentation quality is the same in theory and practice.

Hypothesis 3 Relative quality differences are reflected by differences in absolute quality ratings.

As both corpora described in Section 2 are based on the UKPConvArg1 corpus and thus share many arguments, we can test the hypotheses empirically.

3.2 Correlations of Dimensions and Reasons

For Hypotheses 1 and 2, we consider all 736 pairs of arguments from Habernal and Gurevych (2016a) where both have been annotated by Wachsmuth et al. (2017a). For each pair \((A, B)\) with \(A\) being more convincing than \(B\), we check whether the ratings of \(A\) and \(B\) for each dimension (averaged over all annotators) show a concordant difference (i.e., a higher rating for \(A\)), a disconcordant difference (lower), or a tie. This way, we can correlate each dimension with all reason labels in Table 2 including Conv. In particular, we compute Kendall’s \( \tau \) based on all argument pairs given for each label.2

Table 3 presents all \( \tau \)-values. The phrasing of a reason can be assumed to indicate a clear quality difference—this is underlined by the generally high correlations. Analyzing the single values, we find much evidence for Hypothesis 1: Most notably, label 5-1 perfectly correlates with global acceptability, fitting the intuition that abuse is not acceptable. The high \( \tau \)'s of 8-5 (more credible) for local acceptability (.73) and of 9-4 (well thought through) for cogency (.75) confirm the match assumed in Section 1. Also, the values of 5-3 (unclear) for clarity (.91) and of 7-2 (non-sense) for reasonableness (.94) as well as the weaker correlation of 8-4 (objective) for emotional appeal (.35) makes sense.

Only the comparably low \( \tau \) of 6-1 (no credible evidence) for local acceptability (.49) and credibility (.52) seem really unexpected. Besides, the descriptions of 6-2 and 6-3 sound like local but cor-

1Source code and annotated data: http://www.arguana.com

2Lacking better options, we ignore pairs where a label is not given: It is indistinguishable whether the associated reason does not hold, has not been given, or is just not included in the corpus. Thus, \( \tau \) is more “boosted” the fewer pairs exist for a label and, thus, its values are not fully comparable across labels. Notice, though, that Conv exists for all pairs. So, the values of Conv suggest the magnitude of \( \tau \) without boosting.
As each reason refers to one argument of a pair, this reveals whether the labels, although meant to signal relative differences most: The mean ratings of “negative labels” (5-1 to 7-4) and “positive” ones (8-1 to 9-4). For all dimensions, the maximum and minimum value are higher for the positive than for the negative labels—a clear support of Hypothesis 3.\(^3\) Also, Table 4 reveals which reasons predict absolute differences most: The mean ratings of 7-3 (off-topic) are very low, indicating a strong negative impact, while 6-3 (irrelevant reasons) still shows rather high values. Vice versa, especially 8-5 (more credible) and 9-4 (well thought through) are reflected in high ratings, whereas 9-2 (sticks to topic) does not have much positive impact.

### 3.3 Absolute Ratings for Relative Differences

The correlations found imply that the relative quality differences captured are reflected in absolute differences. For explicitness, we computed the mean rating for each quality dimension of all arguments from Wachsmuth et al. (2017a) with a particular reason label from Habernal and Gurevych (2016a). As each reason refers to one argument of a pair, this reveals whether the labels, although meant to signal relative differences, indicate absolute ratings.

Table 4 compares the mean ratings of “negative labels” (5-1 to 7-4) and “positive” ones (8-1 to 9-4). For all dimensions, the maximum and minimum value are higher for the positive than for the negative labels—a clear support of Hypothesis 3.\(^3\) Also, Table 4 reveals which reasons predict absolute differences most: The mean ratings of 7-3 (off-topic) are very low, indicating a strong negative impact, while 6-3 (irrelevant reasons) still shows rather high values. Vice versa, especially 8-5 (more credible) and 9-4 (well thought through) are reflected in high ratings, whereas 9-2 (sticks to topic) does not have much positive impact.

### 4.1 Absolute Quality Ratings by the Crowd

We emulated the expert annotation process carried out by Wachsmuth et al. (2017a) on CrowdFlower in order to evaluate whether lay annotators suffice for a theory-based quality assessment. In particular, we asked the crowd to rate the same 304 arguments as the experts for all 15 given quality dimensions with scores from 1 to 3 (or choose “cannot judge”). Each argument was rated 10 times at an offered price of $0.10 for each rating (102 annotators in total). Given the crowd ratings, we then performed two comparisons as detailed in the following.

### 4.2 Agreement of the Crowd with Experts

First, we checked to what extent lay annotators and experts agree in terms of Krippendorff’s \(\alpha\). On one hand, we compared the mean of all 10 crowd ratings to the mean of the three ratings of Wachsmuth et al. (2017a). On the other hand, we estimated a reliable rating from the crowd ratings using MACE (Hovy et al., 2013) and compared it to the experts.

<table>
<thead>
<tr>
<th>Polarity</th>
<th>Label</th>
<th>Cog</th>
<th>LA</th>
<th>LR</th>
<th>LS</th>
<th>Eff</th>
<th>Cre</th>
<th>Emo</th>
<th>Cla</th>
<th>App</th>
<th>Arr</th>
<th>Rea</th>
<th>GA</th>
<th>GR</th>
<th>GS</th>
<th>OQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>properties of</td>
<td>5-1</td>
<td>1.30</td>
<td>1.44</td>
<td>1.77</td>
<td>1.29</td>
<td>1.26</td>
<td>1.46</td>
<td>1.64</td>
<td>1.84</td>
<td>1.62</td>
<td>1.55</td>
<td>1.34</td>
<td>1.45</td>
<td>1.65</td>
<td>1.19</td>
<td>1.29</td>
</tr>
<tr>
<td>Argument B</td>
<td>5-2</td>
<td>1.51</td>
<td>1.73</td>
<td>1.97</td>
<td>1.39</td>
<td>1.41</td>
<td>1.66</td>
<td>1.82</td>
<td>1.96</td>
<td>1.89</td>
<td>1.72</td>
<td>1.55</td>
<td>1.72</td>
<td>1.74</td>
<td>1.21</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>5-3</td>
<td>1.46</td>
<td>1.78</td>
<td>2.06</td>
<td>1.43</td>
<td>1.39</td>
<td>1.63</td>
<td>1.96</td>
<td>1.87</td>
<td>2.04</td>
<td>1.65</td>
<td>1.63</td>
<td>1.85</td>
<td>1.76</td>
<td>1.28</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>6-1</td>
<td>1.54</td>
<td>1.87</td>
<td>2.22</td>
<td>1.43</td>
<td>1.44</td>
<td>1.72</td>
<td>1.85</td>
<td>2.15</td>
<td>2.12</td>
<td>1.79</td>
<td>1.62</td>
<td>1.89</td>
<td>1.89</td>
<td>1.27</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td>6-2</td>
<td>1.30</td>
<td>1.52</td>
<td>1.88</td>
<td>1.27</td>
<td>1.21</td>
<td>1.52</td>
<td>1.85</td>
<td>1.94</td>
<td>1.88</td>
<td>1.67</td>
<td>1.36</td>
<td>1.61</td>
<td>1.55</td>
<td>1.15</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>6-3</td>
<td>1.60</td>
<td>1.85</td>
<td>2.23</td>
<td>1.52</td>
<td>1.52</td>
<td>1.65</td>
<td>1.79</td>
<td>2.00</td>
<td>2.15</td>
<td>1.92</td>
<td>1.63</td>
<td>1.85</td>
<td>2.00</td>
<td>1.40</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>7-1</td>
<td>1.43</td>
<td>1.74</td>
<td>1.97</td>
<td>1.33</td>
<td>1.34</td>
<td>1.60</td>
<td>1.82</td>
<td>1.95</td>
<td>1.89</td>
<td>1.72</td>
<td>1.48</td>
<td>1.71</td>
<td>1.68</td>
<td>1.22</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td>7-2</td>
<td>1.45</td>
<td>1.68</td>
<td>1.97</td>
<td>1.41</td>
<td>1.39</td>
<td>1.53</td>
<td>1.86</td>
<td>1.84</td>
<td>1.95</td>
<td>1.67</td>
<td>1.53</td>
<td>1.68</td>
<td>1.70</td>
<td>1.25</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>7-3</td>
<td>1.20</td>
<td>1.47</td>
<td>1.60</td>
<td>1.10</td>
<td>1.17</td>
<td>1.47</td>
<td>1.60</td>
<td>1.70</td>
<td>1.80</td>
<td>1.40</td>
<td>1.20</td>
<td>1.40</td>
<td>1.30</td>
<td>1.07</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>7-4</td>
<td>1.43</td>
<td>1.71</td>
<td>2.02</td>
<td>1.37</td>
<td>1.34</td>
<td>1.71</td>
<td>1.79</td>
<td>1.95</td>
<td>1.97</td>
<td>1.65</td>
<td>1.55</td>
<td>1.75</td>
<td>1.75</td>
<td>1.23</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>8-1</td>
<td>1.56</td>
<td>1.89</td>
<td>2.20</td>
<td>1.46</td>
<td>1.48</td>
<td>1.71</td>
<td>1.88</td>
<td>2.05</td>
<td>2.07</td>
<td>1.79</td>
<td>1.65</td>
<td>1.88</td>
<td>1.92</td>
<td>1.30</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>8-2</td>
<td>1.65</td>
<td>1.97</td>
<td>2.27</td>
<td>1.53</td>
<td>1.61</td>
<td>1.73</td>
<td>1.86</td>
<td>2.12</td>
<td>2.14</td>
<td>1.76</td>
<td>1.73</td>
<td>1.92</td>
<td>1.96</td>
<td>1.37</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>8-3</td>
<td>1.69</td>
<td>2.07</td>
<td>2.39</td>
<td>1.58</td>
<td>1.60</td>
<td>1.81</td>
<td>1.98</td>
<td>2.19</td>
<td>2.25</td>
<td>1.99</td>
<td>1.82</td>
<td>2.04</td>
<td>2.11</td>
<td>1.38</td>
<td>1.75</td>
</tr>
<tr>
<td></td>
<td>8-4</td>
<td>1.54</td>
<td>1.86</td>
<td>2.22</td>
<td>1.49</td>
<td>1.43</td>
<td>1.67</td>
<td>1.84</td>
<td>2.09</td>
<td>2.03</td>
<td>1.74</td>
<td>1.63</td>
<td>1.85</td>
<td>1.92</td>
<td>1.30</td>
<td>1.54</td>
</tr>
<tr>
<td></td>
<td>8-5</td>
<td>1.56</td>
<td>1.76</td>
<td>2.22</td>
<td>1.45</td>
<td>1.49</td>
<td>1.58</td>
<td>1.98</td>
<td>2.02</td>
<td>2.00</td>
<td>1.74</td>
<td>1.62</td>
<td>1.81</td>
<td>1.84</td>
<td>1.28</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td>9-1</td>
<td>1.55</td>
<td>1.78</td>
<td>2.31</td>
<td>1.42</td>
<td>1.49</td>
<td>1.68</td>
<td>2.01</td>
<td>2.18</td>
<td>2.10</td>
<td>1.79</td>
<td>1.63</td>
<td>1.83</td>
<td>1.97</td>
<td>1.27</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>9-2</td>
<td>1.58</td>
<td>1.99</td>
<td>2.32</td>
<td>1.64</td>
<td>1.68</td>
<td>1.81</td>
<td>1.99</td>
<td>2.17</td>
<td>2.19</td>
<td>1.93</td>
<td>1.86</td>
<td>2.05</td>
<td>2.09</td>
<td>1.44</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Table 4: The mean rating for each quality dimension of those arguments from Wachsmuth et al. (2017a) given for each reason label (Habernal and Gurevych, 2016a). The bottom rows show that the minimum maximum mean ratings are consistently higher for the positive properties than for the negative properties.

\(^{3}\)While the differences seem not very large, this is expected, as in many argument pairs from Habernal and Gurevych (2016a) both arguments are strong or weak respectively.
Table 5: Mean and MACE Krippendorff's \( \alpha \) agreement between (a) the crowd and the experts, (b) two independent crowd groups and the experts, (c) group 1 and the experts, and (d) group 2 and the experts.

<table>
<thead>
<tr>
<th>Quality Dimension</th>
<th>Mean</th>
<th>MACE</th>
<th>Mean</th>
<th>MACE</th>
<th>Mean</th>
<th>MACE</th>
<th>Mean</th>
<th>MACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cog Cogency</td>
<td>.27</td>
<td>.38</td>
<td>.24</td>
<td>.29</td>
<td>.38</td>
<td>.37</td>
<td>.05</td>
<td>.27</td>
</tr>
<tr>
<td>LA Local acceptability</td>
<td>.49</td>
<td>.35</td>
<td>.37</td>
<td>.27</td>
<td>.49</td>
<td>.33</td>
<td>.30</td>
<td>.25</td>
</tr>
<tr>
<td>LR Local relevance</td>
<td>.42</td>
<td>.39</td>
<td>.33</td>
<td>.28</td>
<td>.41</td>
<td>.39</td>
<td>.26</td>
<td>.25</td>
</tr>
<tr>
<td>LS Local sufficiency</td>
<td>.18</td>
<td>.31</td>
<td>.21</td>
<td>.21</td>
<td>.34</td>
<td>.27</td>
<td>.04</td>
<td>.19</td>
</tr>
<tr>
<td>Eff Effectiveness</td>
<td>.13</td>
<td>.31</td>
<td>.19</td>
<td>.20</td>
<td>.27</td>
<td>.28</td>
<td>.06</td>
<td>.20</td>
</tr>
<tr>
<td>Cre Credibility</td>
<td>.41</td>
<td>.27</td>
<td>.31</td>
<td>.20</td>
<td>.43</td>
<td>.23</td>
<td>.22</td>
<td>.19</td>
</tr>
<tr>
<td>Emo Emotional appeal</td>
<td>.45</td>
<td>.23</td>
<td>.32</td>
<td>.13</td>
<td>.41</td>
<td>.20</td>
<td>.25</td>
<td>.10</td>
</tr>
<tr>
<td>Cla Clarity</td>
<td>.42</td>
<td>.28</td>
<td>.33</td>
<td>.23</td>
<td>.39</td>
<td>.27</td>
<td>.29</td>
<td>.20</td>
</tr>
<tr>
<td>App Appropriateness</td>
<td>.54</td>
<td>.26</td>
<td>.40</td>
<td>.20</td>
<td>.48</td>
<td>.24</td>
<td>.43</td>
<td>.17</td>
</tr>
<tr>
<td>Arr Arrangement</td>
<td>.53</td>
<td>.30</td>
<td>.36</td>
<td>.24</td>
<td>.49</td>
<td>.27</td>
<td>.35</td>
<td>.24</td>
</tr>
<tr>
<td>Rea Reasonableness</td>
<td>.33</td>
<td>.40</td>
<td>.27</td>
<td>.31</td>
<td>.42</td>
<td>.40</td>
<td>.09</td>
<td>.29</td>
</tr>
<tr>
<td>GA Global acceptability</td>
<td>.54</td>
<td>.40</td>
<td>.36</td>
<td>.29</td>
<td>.53</td>
<td>.37</td>
<td>.33</td>
<td>.28</td>
</tr>
<tr>
<td>GR Global relevance</td>
<td>.44</td>
<td>.31</td>
<td>.31</td>
<td>.20</td>
<td>.50</td>
<td>.29</td>
<td>.22</td>
<td>.18</td>
</tr>
<tr>
<td>GS Global sufficiency</td>
<td>-.17</td>
<td>.19</td>
<td>.04</td>
<td>.11</td>
<td>.00</td>
<td>.16</td>
<td>-.27</td>
<td>.11</td>
</tr>
<tr>
<td>OQ Overall quality</td>
<td>.43</td>
<td>.43</td>
<td>.38</td>
<td>.33</td>
<td>.43</td>
<td>.40</td>
<td>.28</td>
<td>.33</td>
</tr>
</tbody>
</table>

Table 5(a) presents the results. For the mean ratings, most \( \alpha \)-values are above .40. This is similar to the study of Wachsmuth et al. (2017b), where a range of .27 to .51 is reported, meaning that lay annotators achieve similar agreement to experts. Considering the minimum of mean and MACE, we observe the highest agreement for overall quality (.43)—analog to Wachsmuth et al. (2017b). Also, global sufficiency has the lowest agreement in both cases. In contrast, the experts hardly said “cannot judge” at all, whereas the crowd chose it for about 4% of all ratings (most often for global sufficiency), possibly due to a lack of training. Still, we conclude that the crowd generally handles the theory-based quality assessment almost as well as the experts.

However, the complexity of the assessment is underlined by the generally limited agreement, suggesting that either simplification or stricter guidelines are needed. Regarding simplification, the most common practical reasons of Habernal and Gurevych (2016a) imply what to focus on.

4.3 Reliability of the Crowd Annotations

In the second comparison, we checked how many crowd annotators are needed to compete with the experts. For this purpose, we split the crowd ratings into two independent groups of 5 and treated the mean and MACE of each group as a single rating. We then computed the agreement of both groups and each group individually against the experts.

The \( \alpha \)-values for both groups are listed in Table 5(b). On average, they are a bit lower than those of all 10 crowd annotators in Table 5(a). Hence, five crowd ratings per argument seem not enough for sufficient reliability. Tables 5(c) and 5(d) reveal the reason behind, namely, the results of crowd group 1 and group 2 differ clearly. At the same time, the values in Table 5(c) are close to those in Table 5(a), so 10 ratings might suffice. Moreover, we see that the most stable \( \alpha \)-values in Table 5 are given for overall quality, indicating that the theory indeed helps assessing quality reliably.

5 Conclusion

This paper demonstrates that the theory and practice of assessing argumentation quality can learn from each other. Most reasons for quality differences phrased in practice seem well-represented in the normative view of theory and correlate with absolute quality ratings. In our study, lay annotators had similar agreement on the ratings as experts. Considering that some common reasons are quite vague, the diverse and comprehensive theoretical view of argumentation quality may guide a more insightful assessment. On the other hand, some quality dimensions remain hard to assess and/or to separate in practice, resulting in limited agreement. Simplifying theory along the most important reasons will thus improve its practical applicability.

Acknowledgments

We thank Vinodkumar Prabhakaran and Yonatan Bilu for their ongoing participation in our research on argumentation quality. Also, we acknowledge financial support of the DFG (ArguAna, AIPHES), the Natural Sciences and Engineering Research Council of Canada, and the Volkswagen Foundation (Lichtenberg-Professorship Program).
References


