A Closer Look at Evaluation Measures for Ordinal Quantification

Tetsuya Sakai[1]

[1] *Waseda University, Tokyo, Japan*

Abstract

In his ACL 2021 paper [1], Sakai compared several evaluation measures in the context of Ordinal Quantification (OQ) tasks in terms of system ranking similarity, system ranking consistency (i.e., robustness to the choice of test data), and discriminative power (i.e., ability to find many statistically significant differences). Based on his experimental results, he recommended the use of his RNOD (Root Normalised Order-aware Divergence) measure along with NMD (Normalised Match Distance, i.e., normalised Earth Mover's Distance). The present study follows up on his discriminative power experiments, by taking a much closer look at the statistical significance test results obtained from each evaluation measure. Our new analyses show that (1) RNOD is the overall winner among the OQ measures in terms of pooled discriminative power (i.e., discriminative power across multiple data sets); (2) NMD behaves noticeably differently from RNOD and from measures that cannot handle ordinal classes; (3) NMD tends to favour a popularity-based baseline (which accesses the gold distributions) over a uniform-distribution baseline, thus contradicting the other measures in terms of statistical significance. As both RNOD and NMD have their merits, we recommend the organisers of OQ tasks to use both of them to evaluate the systems from multiple angles.

Keywords

evaluation, evaluation measures, distributions, ordinal classes, ordinal quantification, prevalence estimation

1. Introduction

Quantification (or *prevalence estimation*) tasks are highly practical [2, 3, 4]. While *classification evaluation* deals with a confusion matrix whose rows and columns represent gold and estimated classes, *quantification evaluation* compares a gold distribution over the classes with an estimated distribution. Put another way, while classification cares about exactly which of given N instances (with masked gold labels) are classified into each class (represented in the cells of the confusion matrix), quantification only cares about *how many* of the N instances are classified into each class. In general, a quantification task involves n test cases; each test case has N instances, and N can vary across cases. Hence a quantification evaluation measure computes a score for each test case by comparing the gold and estimated distributions, and the measure score can be averaged across the n cases. Hence statistical significance tests can be applied to compare the systems.

In the present study, if a task involves the comparison of an estimated probability mass function with a gold probability mass function for each test case, we regard it as a quantification task, regardless of what the exact input to the estimation system is. (Note that a frequency distribution of N instances can easily converted to a probability mass function.) In particular, we define *Ordinal Quantification* (OQ) as a task that requires systems to estimate a probability mass function over *ordinal* classes for each of n test cases.[1] Examples of OQ tasks defined in this way include the following.

SemEval 2016/2017 Task 4 Subtask E Given a set of N tweets about a topic, estimate the distribution of the tweets over five classes: *highly negative, negative, neutral, positive, highly positive* [5, 6].

Dialogue Breakdown Detection Challenge For each system utterance within a human-machine dialogue, estimate the distribution of gold labels given by N annotators, where the possible labels are *NB (not a breakdown), PB (possible breakdown), B (breakdown)* [7].

NTCIR Dialogue Quality Subtasks For each customer-helpdesk dialogue, estimate the distribution of dialogue quality ratings given by N annotators, where the possible ratings are -2, -1, *0, 1, 2* [8, 9].

To evaluate systems (or *runs* [22]) submitted to OQ tasks, evaluation measures that can handle ordinal classes should be used. More specifically, "nominal quantification" measures such as Mean Absolute Error (MAE), (Root) Mean Squared Error ((R)MSE), and Jensen-Shannon Divergence (JSD) are not adequate, as they are based on simple averaging/summing across classes [1].

First International Workshop on Learning to Quantify: Methods and Applications (LQ 2021), November 1, 2021, Gold Coast, Queensland, Australia

tetsuyasakai@acm.org (T. Sakai)
0000-0002-6720-963X (T. Sakai)

[1] Interval classes are also ordinal by definition.

Table 1
Quantification measures considered in this study.

Ordinal quantification measures	
NMD	Normalised Match Distance (Normalised Earth Mover's Distance)
RNOD	Root Normalised Order-aware Divergence
RSNOD	Root Symmetric Normalised Order-aware Divergence (symmetric version of RNOD)
Nominal quantification measures	
NVD	Normalised Variational Distance (essentially Mean Absolute Error)
RNSS	Root Normalised Sum of Squares (essentially Root Mean Squared Error)
JSD	Jensen-Shannon Divergence (symmetric version of Kullback-Leibler Divergence)

To see why, consider a gold distribution for the aforementioned SemEval task, where all N tweets for a topic are in the *highly positive* class; consider a system which puts all N tweets in *highly negative* (i.e., an utter failure), and consider another which puts all N tweets in *positive*. It is clear that the above measures rate both systems as utter failures.

To the best of our knowledge, only two families of measures are known to be suitable for evaluating OQ systems: the *Earth Mover's Distance* family [10, 11, 12], which is based on cumulative distributions of the gold and estimated distributions, and Sakai's *Order-aware Divergence* family, proposed in 2017-2018 [13, 14]. Recently, Sakai [1] reported on extensive experiments for comparing the above two families as well as nominal quantification measures in the context of evaluating OQ systems submitted to the SemEval and NTCIR tasks. His recommendation was to use *Root Normalised Order-aware Divergence* (RNOD) as the primary measure, and *Normalised Match Distance* (NMD) as the secondary measure, where NMD is simply a normalised version of Earth Mover's Distance [5]. In that study, RNOD was preferred over NMD because it was the overall winner when looked across the data sets in terms of *system ranking consistency* (i.e., the ability to provide stable system rankings regardless of the choice of test data) [15] and *discriminative power* (i.e., the ability to obtain many statistically significance differences under the same experimental condition) [16, 17].

The present study follows up on Sakai's experiments [1], by taking a much closer look at the statistical significance test results obtained from each evaluation measure. We also leverage additional sets of OQ data from NTCIR that were not previously used. Our new analyses show that (1) RNOD is the overall winner among the OQ measures in terms of *pooled discriminative power* (i.e., discriminative power across multiple data sets); (2) NMD behaves noticeably differently from RNOD and from measures that cannot handle ordinal classes; (3) NMD tends to favour a popularity-based baseline (which accesses the gold distributions) over a uniform-distribution baseline, thus contradicting the other measures in terms of statistical significance. As both RNOD and NMD have their merits, we recommend the organisers of OQ tasks to use both of them to evaluate the systems from multiple angles.

2. Related Work

The aforementioned OQ tasks of SemEval (2016/2017 Task 4 Subtask E) [5, 6] used Earth Mover's Distance (EMD) as the evaluation meaure, remarking that "*EMD is currently the only known measure for ordinal quantification.*" Their EMD is the same as *Match Distance* [14, 10], and the present study uses its normalised version, called Normalised Match Distance (NMD) [14, 1]. NMD has been used as one of the evaluation measures for evaluating the aforementioned OQ tasks of NTCIR (Dialogue Quality) [8, 9].

In 2017, Sakai [13] proposed Order-aware Divergence (OD), Normalised OD (NOD), and Symmetric Normalised OD (SNOD) for OQ tasks, by explicitly incorporating the notion of "distance" between classes. Subsequently, Sakai [14] proposed *Root* Normalised OD (RNOD) and *Root* Symmetric Normalised OD (RSNOD), as the computation of OD involves sums of squares. The OQ tasks of NTCIR have used RSNOD along with NMD [8, 9]. Sakai's recent recommendation for OQ tasks [1] is to use RNOD as the primary measure and NMD as the secondary measure, for the reasons discussed in Section 1.

Although the aforementioned Dialogue Breakdown Detection Challenge (DBDC) [7] is an OQ task, the official evaluation measures used there for comparing two distributions were MSE and JSD, which cannot consider the ordinal nature of classes (i.e., nominal quantification measures). Subsequently, the organisers of DBDC used their Japanese and English DBDC task data to compare these official DBDC measures with NMD and RSNOD in terms of system ranking consistency and discriminative power; they reported that RSNOD was the overall winner [18].

3. Quantification Measures

Table 1 provides a brief qualitative summary of the measures considered in this study. Due to lack of space, we refer the reader to Sakai [1] for the definitions of nominal quantification measures; here, we define only the ordinal quantification measures.

Let C denote the set of ordinal classes, represented by consecutive integers for convenience. Let p_i denote the estimated probability for Class i, so that $\sum_{i \in C} p_i = 1$. Similarly, let p^*_i denote the gold probability. We also denote the entire probability mass functions by p and p^*, respectively. Let $cp_i = \sum_{k \leq i} p_k$, and $cp^*_i = \sum_{k \leq i} p^*_k$. NMD is given by [14]:

$$NMD(p, p^*) = \frac{\sum_{i \in C} |cp_i - cp^*_i|}{|C| - 1} . \tag{1}$$

We now define R(S)NOD. First, let the *Distance-Weighted sum of squares* for Class i be:

$$DW_i = \sum_{j \in C} \delta_{ij} (p_j - p^*_j)^2 , \quad \delta_{ij} = |i - j| . \tag{2}$$

DW_i was designed to quantify the overall error from the viewpoint of a particular gold class i: it tries to measure how much of its probability p^*_i has been misallocated to other classes $j \in C (j \neq i)$, by assuming that the difference between p_j and p^*_j is directly caused by a misallocation of part of p^*_i; the weight δ_{ij} is designed to penalise the misallocation based on the distance between the ordinal classes. Note that the δ_{ij} in Eq. 2 assumes equidistance; we shall discuss an alternative in Section 6.

Let $C^* = \{i \in C | p^*_i > 0\}$. That is, $C^* (\subseteq C)$ is the set of classes with a non-zero gold probability. *Order-aware Divergence* is defined as:

$$OD(p \parallel p^*) = \frac{1}{|C^*|} \sum_{i \in C^*} DW_i , \tag{3}$$

with its symmetric version:

$$SOD(p, p^*) = \frac{OD(p \parallel p^*) + OD(p^* \parallel p)}{2} . \tag{4}$$

RNOD and RSNOD are defined as:

$$RNOD(p \parallel p^*) = \sqrt{\frac{OD(p \parallel p^*)}{|C| - 1}} , \tag{5}$$

$$RSNOD(p, p^*) = \sqrt{\frac{SOD(p, p^*)}{|C| - 1}} . \tag{6}$$

Note that Eq. 3 averages over C^* rather than C because of what DW_i is meant to represent, as discussed above. However, it is also possible to define a variant of OD as follows; let us call it ADW (Average DW_i):

$$ADW(p, p^*) = \frac{1}{|C|} \sum_{i \in C} DW_i . \tag{7}$$

From Eqs. 2 and 7, it is clear that ADW is symmetric.[2] A root-normalised measure based on ADW, which we call RNADW (Root Normalised ADW) can also be defined:

$$RNADW(p, p^*) = \sqrt{\frac{ADW(p, p^*)}{|C| - 1}} . \tag{8}$$

We will evaluate this variant in our future work.

When there are only two classes ($|C| = 2$) and therefore the distinction between nominal and ordinal classes becomes unnecessary, it can be shown that $NMD(p, p^*) = RNOD(p \parallel p^*) = RNOD(p^* \parallel p) (= RSNOD(p, p^*))$. See the Appendix for a proof.

OD-based measures tend to emphasise errors near either end of the ordinal scale, as the following example illustrates. Consider a situation with $|C| = 4$ ordinal classes and a uniform gold distribution: $p^*_i = 0.25 (i = 1, \ldots, 4)$. If we compare System A which returns $p_1 = p_4 = 0.25, p_2 = 0.35, p_3 = 0.15$ and System B which retruns $p_1 = p_2 = 0.25, p_3 = 0.35, p_4 = 0.15$, then from Eq. 2, $DW_1 = DW_4 = 0.03, DW_2 = DW_3 = 0.01$ and therefore $OD = 0.020$ for System A; whereas, $DW_1 = 0.05, DW_2 = 0.03, DW_3 = DW_4 = 0.01$ and therfore $OD = 0.025$ for System B. Hence System A is considered slightly better. On the other hand, it can easily be verified that A and B are considered equally effective in terms of NMD. It should be noted that this difference does not say which measure is "correct" as an OQ evaluation measure, as both measures take the ordinal nature of the classes into account. (Similarly, we cannot say whether (say) JSD is superior to NVD for a nominal quantification task just because they differ.)

4. Data

Table 2 provides an overview of the eight OQ task data sets that we used for our analysis. The three STC-3 (Short Text Conversation 3) data sets [8] were not used in Sakai [1], but the specifications of the Dialogue Quality (DQ) subtask at STC-3 are identical to those of DialEval-1 (Dialogue Evaluation 1) [9]. As can be seen, all data sets come with five ordinal classes. For the two SemEval data sets, the classes are tweet polarities, namely, *highly negative, negative, neutral, positive, highly positive* [5, 6]. For the six NTCIR data sets, the classes are five-point scale dialogue quality ratings (-2 through 2) based on three different viewpoints, namely, *A-score* (task accomplishment), *E-score* (dialogue effectiveness), and *S-score* (customer satisfaction) [8, 9]. Hence, for example, DialEval-1 DQ-A is the data set containing the gold and estimated probability distributions for the A-score estimation "sub-subtask" of the NTCIR-15 DialEval-1 task. The NTCIR dialogue data were provided in both Chinese and English (manually translated from the original Chinese text) to the participants, and the participants were allowed to submit Chinese and/or English runs. On the other hand,

[2] Similarly, it is clear from Eqs. 2 and 3 that $C^* = C$ (i.e., there is no gold probability that is zero) is a sufficient condition for OD to be symmetric [13]. Another sufficient condition for guaranteeing the symmetry of OD is: $|C^*| = 1$ *and* $|\{p_i \in C \mid p_i > 0\}| = 1$ (i.e., both the gold and estimated distributions have exactly one positive probability).

Table 2
Eight data sets used in our experiments (C: Chinese; E: English; C+E: runs for both languages combined).

Short name in this paper	Evaluation venue	Task (Subtask)	Task Type	Language	#Ordinal classes	Test data sample size	#Runs used
Sem16T4E	SemEval-2016	Task 4 (Subtask E)	OQ	E	5	100	12
Sem17T4E	SemEval-2017	Task 4 (Subtask E)	OQ	E	5	125	14
STC-3 DQ-{A, E, S}	NTCIR-14 (2019)	Short Text Conversation 3 (Dialogue Quality)	OQ	C+E	5	390	19 (10+9)
DialEval-1 DQ-{A, E, S}	NTCIR-15 (2020)	Dialogue Evaluation 1 (Dialogue Quality)	OQ	C+E	5	300	22 (13+9)

Table 3
Discriminative powers of ordinal quantification measures (NMD, RSNOD, RNOD) and nominal quantification measures (NVD, RNSS, JSD) at significance level $\alpha = 0.05$. "#all" is the number of system pairs compared; "#sig" is the number of systems pairs with a statistically significant difference according to a randomised Tukey HSD test with $B = 5,000$ trials. Percentages over or equal to 50 are shown in **bold**; Those below 40 are underlined.

	NMD			RSNOD			RNOD		
	S_D	T_D	%	S_D	T_D	%	S_D	T_D	%
Sem16T4E	38	66	**57.6**	32	66	48.5	35	66	**53.0**
Sem17T4E	48	91	**52.7**	40	91	44.0	35	91	38.5
STC-3 DQ-A	71	171	41.5	67	171	39.2	68	171	39.8
STC-3 DQ-E	68	171	39.8	72	171	42.1	66	171	38.6
STC-3 DQ-S	65	171	38.0	66	171	38.6	61	171	35.7
DialEval-1 DQ-A	84	231	36.4	119	231	**51.5**	133	231	**57.6**
DialEval-1 DQ-E	116	231	**50.2**	116	231	**50.2**	117	231	**50.6**
DialEval-1 DQ-S	82	231	35.5	115	231	49.8	125	231	**54.1**
POOLED	572	1363	42.0	627	1363	46.0	640	1363	47.0

	NVD			RNSS			JSD		
	S_D	T_D	%	S_D	T_D	%	S_D	T_D	%
Sem16T4E	31	66	47.0	37	66	**56.1**	32	66	48.5
Sem17T4E	40	91	44.0	37	91	40.7	35	91	38.5
STC-3 DQ-A	68	171	39.8	67	171	39.2	64	171	37.4
STC-3 DQ-E	66	171	38.6	65	171	38.0	64	171	37.4
STC-3 DQ-S	61	171	35.7	65	171	38.0	64	171	37.4
DialEval-1 DQ-A	138	231	**59.7**	113	231	48.9	135	231	**58.4**
DialEval-1 DQ-E	116	231	**50.2**	115	231	49.8	115	231	49.8
DialEval-1 DQ-S	120	231	**51.9**	129	231	**55.8**	127	231	**55.0**
POOLED	640	1363	47.0	628	1363	46.1	636	1363	46.7

the gold distributions were constructed solely based on the original Chinese dialogues. Hence, both Chinese and English runs are evaluated using the same gold distributions. The gold distributions of the STC-3 and DialEval-1 data were constructed based on votes from 19 and 20 assessors for each dialogue, respectively [8].

The NTCIR data sets are larger than the SemEval data sets both in terms of the test data sample size n and in terms of the number of runs to be evaluated. Hence our results with the NTCIR data may be more reliable, especially regarding statistical significance test results.

5. Analysis

5.1. Pooled Discriminative Power

Sakai [1] presented *discriminative power curves* [16] for NMD, R(S)NOD, NVD, RNS, and JSD using the SemEval and DialEval-1 data sets. Given a data set with submitted runs, a discriminative power curve is obtained by obtaining a p-value for every system pair (using a randomised Tukey HSD test with $B = 5000$ trials [19]) and sorting them in descending order. Curves that are closer to the origin represent discriminative measures, i.e., those that can give us confident conclusions from experiments. A highly discriminative measure is not necessarily "correct," but we do want measures to be discriminative to some extent; otherwise, we will not be able to conclude anything from experiments [20].

Here, we revisit Sakai's results by focusing on the commonly-used significance level of $\alpha = 0.05$, to view and summarise the discriminative power results in a more quantitative manner. More specifically, for data set D, let $DP_D = S_D/T_D$, where T_D is the **t**otal number of system pairs ($T_D = m_D(m_D - 1)/2$ if there are m_D systems) and $S_D(\leq T_D)$ is the number of those found to be statistically **s**ignificantly different at α. To provide a quantitative summary of discriminative power results over a set $\mathcal{D} = \{D\}$ of data sets, we define *pooled* discriminative power as follows:

$$PDP = \sum_{D \in \mathcal{D}} S_D / \sum_{D \in \mathcal{D}} T_D \ . \tag{9}$$

We also report on additional results with the three STC-3 data sets; these were not discussed in Sakai [1].

Table 3 shows the individual and pooled discriminative power results for each of our eight OQ data sets. For example, for NMD, the discriminative power with Sem16T4E is $38/66 = 57.6\%$ and higher than the other measures, but as it suffers with the NTCIR data, the pooled discriminative power is only 42.0%. In particular, note that NMD performs very poorly with DialEval-1 DQ-A and DQ-S data sets: with DialEval-1 DQ-A, NMD finds only 84 statistically significant differences at $\alpha = 0.05$, while RSNOD, RNOD, NVD, RNSS, and JSD find as many as 119, 133, 138, 113, and 135, respectively; similarly, with DialEval DQ-S, NMD finds only 82 statistically significant differences at $\alpha = 0.05$, while RSNOD, RNOD, NVD, RNSS, and JSD find as many as 115, 125, 120, 129, and 127, respectively. This apparent breakdown of NMD for these two data sets was also visualised in the discriminative curves of Sakai [1, Figure 2].

Our findings in terms of pooled discriminative power are:

- The most discriminative measures are RNOD and NVD (but recall that NVD is a nominal quantification measure).
- RNOD outperforms NMD;
- RSNOD slightly underperforms RNOD, suggesting that making the measure symmetric is not beneficial [1].

5.2. Significance Overlaps and Contradictions

Discriminative power only considers how many significant differences each measure manages to obtain; it does not examine which measures agree or disagree with each other in terms of significance test results. This section addresses exactly this question.

Table 4 breaks down the number of significant differences (S_D) shown in Table 3 by comparing the results of every pair of measures. More specifically, we present *Statistical Significance Overlaps* (SSO's), defined as $SSO = b/(a + b + c)$, where a is the number of significant differences found with the first measure only, b is the number of significant differences found with both measures, and c is the number of significant differences found with the second measure only. That is, the S_D for the first measure is $a + b$, and that for the second measure is $b + c$.

If the SSO for a pair of evaluation measures is high, that means that the two measures tend to give us similar conclusions as to which system pairs are statistically significantly different. However, it can be observed that SSO is not always high, as underlined in Table 4. In particular, note that the SSOs of NMD with other measures are particularly low for DialEval-1 DQ-A and DQ-S (Parts (f) and (h) of the table). In Section 5.1, we have pointed out that NMD performs very poorly with these two data sets in terms of discriminative power. The discriminative power results alone could mean two situations: (i) NMD manages to find only a subset of the significant differences found by the other measures; *or* (ii) NMD finds significant differences outside those found by the other measures, *and* the differences found are relatively few. Table 4(f) and (h) reveals that the truth is Situation (ii). For example, from Table 4(h), we can see that the SSO between NMD and RNOD is only 52.2% (with only 71 differences found significant by both measures), and that NMD found as many as 11 significant differences that were not considered significant by RNOD. Similarly, the SSO between NMD and NVD is only 48.5% (with only 66 differences found significant by both measures), and NMD found as many as 16 significant differences that were not considered significant by NVD. This outlier tendency of NMD is consistent with Sakai's observation regarding system ranking similarity [1, Table 5].

The above analysis examined the overlaps of significantly different system pairs based on two-sided tests. However, the overlaps in fact contain a few *contradictions*: a statistical significance contradiction occurs when one measure says "System A statistically significantly outperforms System B" while another says "System B statistically significantly outperforms System A." Which system outperforms another is determined by the mean scores of A and B (smaller the better in our case). Although such situations are very rare, we have found them useful for understanding the properties of the measures, as discussed below.

Table 5 shows the number of contradictions, which can be used together with Table 4. (There were no contra-

Table 4

Statistical Significance Overlap (SSO) for every pair of measures ($\alpha = 0.05$). SSO$= b/(a + b + c)$ (in percentages) where a: #significant with Measure 1 only; b: #significant with both Measures; c: #significant with Measure 2 only. a, b, c are shown as "(a/b/c)." Percentages over or equal to 90 are shown in **bold**; those below 70 are underlined.

(a) Sem16T4E	RSNOD	RNOD	NVD	RNSS	JSD
NMD	70.7 (9/29/3)	65.9 (9/29/6)	76.9 (8/30/1)	59.6 (10/28/9)	79.5 (7/31/1)
RSNOD	-	86.1 (1/31/4)	85.3 (3/29/2)	76.9 (2/30/7)	73.0 (5/27/5)
RNOD	-	-	78.4 (6/29/2)	89.5 (1/34/3)	67.5 (8/27/5)
NVD	-	-	-	70.0 (3/28/9)	85.3 (2/29/3)
RNSS	-	-	-	-	60.5 (11/26/6)
(b) Sem17T4E	RSNOD	RNOD	NVD	RNSS	JSD
NMD	76.0 (10/38/2)	66.0 (15/33/2)	76.0 (10/38/2)	63.5 (15/33/4)	72.9 (13/35/0)
RSNOD	-	87.5 (5/35/0)	**90.5 (2/38/2)**	79.1 (6/34/3)	87.5 (5/35/0)
RNOD	-	-	87.5 (0/35/5)	89.5 (1/34/3)	84.2 (3/32/3)
NVD	-	-	-	83.3 (5/35/2)	87.5 (5/35/0)
RNSS	-	-	-	-	75.6 (6/31/4)
(c) STC-3 DQ-A	RSNOD	RNOD	NVD	RNSS	JSD
NMD	**91.7** (5/66/1)	**95.8** (3/68/0)	**95.8** (3/68/0)	**94.4** (4/67/0)	**90.1** (7/64/0)
RSNOD	-	**90.1** (3/64/4)	**90.1** (3/64/4)	88.7 (4/63/4)	84.5 (7/60/4)
RNOD	-	-	**100.0** (0/68/0)	**98.5** (1/67/0)	**94.1** (4/64/0)
NVD	-	-	-	**98.5** (1/67/0)	**94.1** (4/64/0)
RNSS	-	-	-	-	**95.5** (3/64/0)
(d) STC-3 DQ-E	RSNOD	RNOD	NVD	RNSS	JSD
NMD	**94.4** (0/68/4)	**97.1** (2/66/0)	**97.1** (2/66/0)	**95.6** (3/65/0)	**94.1** (4/64/0)
RSNOD	-	**91.7** (6/66/0)	**91.7** (6/66/0)	**90.3** (7/65/0)	88.9 (8/64/0)
RNOD	-	-	**100.0** (0/66/0)	**98.5** (1/65/0)	**97.0** (2/64/0)
NVD	-	-	-	**98.5** (1/65/0)	**97.0** (2/64/0)
RNSS	-	-	-	-	**98.5** (1/64/0)
(e) STC-3 DQ-S	RSNOD	RNOD	NVD	RNSS	JSD
NMD	**98.5** (0/65/1)	**93.8** (4/61/0)	**93.8** (4/61/0)	**100.0** (0/65/0)	**98.5** (1/64/0)
RSNOD	-	**92.4** (5/61/0)	**92.4** (5/61/0)	**98.5** (1/65/0)	**97.0** (2/64/0)
RNOD	-	-	**100.0** (0/61/0)	**93.8** (0/61/4)	**92.3** (1/60/4)
NVD	-	-	-	**93.8** (0/61/4)	**92.3** (1/60/4)
RNSS	-	-	-	-	**98.5** (1/64/0)
(f) DialEval-1 DQ-A	RSNOD	RNOD	NVD	RNSS	JSD
NMD	65.0 (4/80/39)	55.0 (7/77/56)	56.3 (4/80/58)	39.7 (28/56/57)	62.2 (0/84/51)
RSNOD	-	81.3 (6/113/20)	82.3 (3/116/22)	65.7 (27/92/21)	77.6 (8/111/24)
RNOD	-	-	**92.2** (3/130/8)	83.6 (21/112/1)	**90.1** (6/127/8)
NVD	-	-	-	76.8 (29/109/4)	**92.3** (7/131/4)
RNSS	-	-	-	-	74.6 (7/106/29)
(g) DialEval-1 DQ-E	RSNOD	RNOD	NVD	RNSS	JSD
NMD	77.1 (15/101/15)	80.6 (12/104/13)	82.7 (11/105/11)	81.9 (12/104/11)	81.9 (12/104/11)
RSNOD	-	**94.2** (3/113/4)	**93.3** (4/112/4)	**92.5** (5/111/4)	**92.5** (5/111/4)
RNOD	-	-	**97.5** (2/115/1)	**98.3** (2/115/0)	**98.3** (2/115/0)
NVD	-	-	-	**99.1** (1/115/0)	**99.1** (1/115/0)
RNSS	-	-	-	-	**100.0** (0/115/0)
(h) DialEval-1 DQ-S	RSNOD	RNOD	NVD	RNSS	JSD
NMD	60.2 (8/74/41)	52.2 (11/71/54)	48.5 (16/66/54)	50.7 (11/71/58)	56.0 (7/75/52)
RSNOD	-	77.8 (10/105/20)	75.4 (14/101/19)	80.7 (6/109/20)	80.6 (7/108/19)
RNOD	-	-	**92.9** (7/118/2)	**91.0** (4/121/8)	**90.9** (5/120/7)
NVD	-	-	-	**91.0** (4/121/8)	**90.9** (5/120/7)
RNSS	-	-	-	-	**91.0** (7/122/5)

Table 5
Statistical Significance Contradictions at $\alpha = 0.05$. (There were no contradictions in the Sem16T4E, Sem17T4E, and STC-3 DQ-E results.)

(I) STC-3 DQ-A	RSNOD	RNOD	NVD	RNSS	JSD
NMD	0	4	4	4	4
RSNOD	-	0	0	0	0
RNOD	-	-	0	0	0
NVD	-	-	-	0	0
RNSS	-	-	-	-	0
(II) STC-3 DQ-S	**RSNOD**	**RNOD**	**NVD**	**RNSS**	**JSD**
NMD	0	0	0	4	4
RSNOD	-	0	0	4	4
RNOD	-	-	0	0	0
NVD	-	-	-	0	0
RNSS	-	-	-	-	0
(III) DialEval-1 DQ-A	**RSNOD**	**RNOD**	**NVD**	**RNSS**	**JSD**
NMD	0	8	6	8	8
RSNOD	-	8	6	8	8
RNOD	-	-	0	0	0
NVD	-	-	-	0	0
RNSS	-	-	-	-	0
(IV) DialEval-1 DQ-E	**RSNOD**	**RNOD**	**NVD**	**RNSS**	**JSD**
NMD	0	4	4	4	4
RSNOD	-	0	0	0	0
RNOD	-	-	0	0	0
NVD	-	-	-	0	0
RNSS	-	-	-	-	0
(V) DialEval-1 DQ-S	**RSNOD**	**RNOD**	**NVD**	**RNSS**	**JSD**
NMD	0	4	0	8	8
RSNOD	-	4	0	8	8
RNOD	-	-	0	0	0
NVD	-	-	-	0	0
RNSS	-	-	-	-	0

dictions in the Sem16T4E, Sem17T4E, and STC-3 DQ-E results.) For example, while Table 4(c) shows that NMD and RNOD detected a statistical significance for the same 68 run pairs from STC-3 DQ-A (with an SSO of 95.8%), Table 5(I) shows that 4 of them were actually contradictions. Hence, if we choose to remove the contradictions prior to computing the SSO, it would be $64/(3+64+0) = 95.5\%$. However, "practical" contradictions in the five NTCIR data sets occur less frequently than what Table 5 suggests, as explained below.

Table 6 provides an exhaustive list of the exact run pairs listed as contradictions in Table 5. For example, Table 6(I) reveals that the 4 contradictions mentioned in Table 5(a) are:

- A_BL-popularity-C vs. A_BL-uniform-C
- A_BL-popularity-C vs. A_BL-uniform-E
- A_BL-popularity-E vs. A_BL-uniform-C
- A_BL-popularity-E vs. A_BL-uniform-E

However, these essentially constitute *one* contradiction, as the contents of the Chinese and English runs A_BL-popularity-{C,E} are the same, as are those of A_BL-uniform-{C,E}. These are the *Popularity and Uniform baseline runs* provided by the organisers of the NTCIR tasks, which rely on the following simple strategies.[3]

Popularity Access the gold data, and return an "estimated" distribution where the class that is most frequent in the gold distribution is given a probability of 1, and others are given a 0. Note that this is a type of oracle run.

Uniform Always return a uniform distribution.

[3] The SemEval tasks also had a few baseline runs, including a run that always assigns a probability of 1 to the Positive class [5, 6]. They did not have Popularity and Uniform baselines.

Table 6
Details of contradicting run pairs, breaking down the numbers shown in Table 5. Measure 1 says Run 1 statistically significantly outperforms Run 2 while Measure 2 says Run 1 statistically significantly underperforms Run 2.

(I) STC-3 DQ-A

NMD (Measure 1) contradicts with other measures for the following run pairs:
With RNOD, NVD, RNSS, JSD (Measure 2):
A_BL-popularity-{C,E} and A_BL-uniform-{C,E}

(II) STC-3 DQ-S

NMD and RSNOD (Measure 1) contradict with other measures for the following run pairs:
With RNSS and JSD (Measure 2):
S_BL-popularity-{C,E} S_BL-uniform-{C,E}

(III) DialEval-1 DQ-A
NMD and RSNOD (Measure 1) contradict with other measures for the following run pairs: **With RNOD, RNSS and JSD (Measure 2):** A_BL-popularity-{C,E} and A_BL-uniform-{C,E} A_BL-popularity-{C,E} and A_NKUST-run0-C A_BL-popularity-{C,E} and A_NKUST-run0-E **With NVD:** A_BL-popularity-{C,E} and A_BL-uniform-{C,E} A_BL-popularity-{C,E} and A_NKUST-run0-C
(IV) DialEval-1 DQ-E

NMD (Measure 1) contradicts with other measures for the following run pairs:
With RNOD, NVD, RNSS, and JSD (Measure 2):
E_BL-popularity-{C,E} and E_BL-uniform-{C,E}

(V) DialEval-1 DQ-S
NMD and RSNOD (Measure 1) contradict with other measures for the following run pairs: **With RNOD (Measure 2):** S_BL-popularity-{C,E} and S_NKUST-run0-C S_BL-popularity-{C,E} and S_NKUST-run0-E **With RNSS and JSD (Measure 2):** S_BL-popularity-{C,E} and S_BL-uniform-{C,E} S_BL-popularity-{C,E} and S_NKUST-run0-C S_BL-popularity-{C,E} and S_NKUST-run0-E

Note that the prefix "A_BL" means that the run is a baseline for the A-score estimation subsubtask; similar baseline runs are present in the E-score and S-score estimation subsubtask data from NTCIR, with prefixes "E_BL" and "S_BL."

Sakai [1] kept both the Chinese and English versions of the baselines in his experiments even though their contents are the same, because they have different file names and were listed as distinct runs in the official evaluations [8, 9]. Hence we follow suit. However, it is clear from the above that the contradiction in STC-3 DQ-A is essentially a single instance: while NMD concludes that Popularity statistically significantly outperforms Uniform, RNOD, NVD, RNSS, and JSD concludes the exact opposite. Similarly, Table 6(II) reveals that the 4 contradictions shown in Table 5 also concerns Popularity vs. Uniform: NMD and RSNOD conclude that Popularity statistically significantly outperforms Uniform, while RNSS and JSD conclude the exact opposite. As for the results for DialEval-1 DQ-E (Table 6(IV)), they are essentially identical to those of STC-3 DQ-A (Table 6(I)).

In Table 6(III) and (V), we see non-baseline runs. Note that, for example, A_NKUST-run0-C and A_NKUST-run0-E are actually different runs, unlike the situations with the aforementioned baseline runs. Thus, for example, the 8 contradictions shown in Table 5(III) between NMD/RSNOD and RNOD/RNSS/JSD are essentially for the following 3 cases, as shown in Table 6(III).

- A_BL-popularity-{C,E} vs. A_BL-uniform-{C,E} (4 run combinations)
- A_BL-popularity-{C,E} vs. A_NKUST-run0-C (2 run combinations)
- A_BL-popularity-{C,E} vs. A_NKUST-run0-E (2 run combinations)

In every case, NMD and RSNOD conclude that Popularity statistically significantly outperforms the other run, which is in direct disagreement with the other measures.

We also observe that RSNOD behaves similarly to NMD from Table 5(II), (III), and (V) and the accompanying details in Table 6(II), (III), and (V). Moreover, there are no contradictions between NMD and RSNOD. These results suggest that the properties of RSNOD lie somewhere between NMD and RNOD: this is in line with Sakai's observation regarding the system ranking similarity for the DialEval-1 DQ-A and DQ-S data [1, Table 5]. Put another way, introducing symmetry appears to bring RNOD closer to NMD.

Our findings regarding statistical significance overlap and contradictions can be summarised as follows.

- The sets of significant differences found by NMD are generally *not* subsets of those found by the other, more discriminative measures. NMD behaves markedly differently from other measures regarding which system pairs are statistically significant.
- There are a few contradictions between NMD and four other measures (RNOD, NVD, RNSS, JSD) in terms of significance test results, and all of these contradictions involve a Popularity baseline, which access the gold distributions. NMD tends to rate Popularity higher than Uniform, thus directly contradicting the other measures.
- From the viewpoint of contradictions, RSNOD behaves somewhat similarly to NMD.

5.3. Popularity vs. Uniform baselines

In Section 5.2, we showed that NMD can contradict with RNOD, NVD, RNSS or JSD in terms of statistical significance. In particular, we have seen cases where NMD favours Popularity over Uniform, contrary to the conclusions of the other measures. We find this behaviour of NMD generally intuitive, as Popularity accesses the gold data and utilises that knowledge, while Uniform is noninformative and practically useless. However, note that whether Popularity should actually be rated higher depends on what the gold distribution looks like: for example, if the gold distributions is almost flat, we would like the measure to prefer Uniform over Popularity.

To examine the above tendency of NMD, this section focusses on the comparison between Popularity and Uniform. First, we focus on contradictions regarding Popularity vs. Uniform from the DialEval-1 DQ-A data set, as we found the highest number of conflicts (not limited to Popularity vs. Uniform) in this data set among the five data sets shown in Table 5-6. For each evaluation measure M and for each dialogue, we first compute the *score delta* (e.g. ΔNMD):

$$\Delta M = M(\text{Popularity}) - M(\text{Uniform}) \ , \qquad (10)$$

where $M(\bullet)$ is the score according to measure M for a run's estimated distribution for a particular dialogue. Note that, since these measures give smaller scores to better systems, a negative delta means Popularity is preferred while a positive delta means Uniform is preferred.

Figure 1 shows scatter plots of score deltas by comparing NMD with NVD, RNSS, JSD, and RNOD; Figure 2 shows similar scatterplots by comparing RNOD with NVD, RNSS, and JSD. The two figures are arranged to facilitate comparisons across NMD and RNOD. (To reduce the number of measure combinations, RSNOD is omitted in this analysis.) Within each green box, the number of instances in the 2nd and 4th quadrants (i.e., dialogues for which two measures disagree as to which of Popularity and Uniform is better) is shown, together with a Pearson correlation with a 95%CI. From the figures, we can observe the following.

- The correlations between NMD and the nominal quantification measures (NVD, RNSS, and JSD) are lower compared to those between RNOD and the nominal quantification measures, as the Pearson correlations and the scatterplots show.
- More importantly, NMD disagrees more often with the nominal quantification measures than RNOD does. All of these disagreements of NMD happen in the 2nd quadrant: that is, while NMD says that Popularity outperforms Uniform, the other three measures say otherwise.
- From Figure 1(d), NMD and RNOD disagree for a total of 77 dialogues: for 75 of them, NMD says that Popularity outperforms Uniform while RNOD says otherwise; for the remaining 2 dialogues, NMD says that Popularity *underperforms* Uniform while RNOD says otherwise.

In summary, for 62-78 dialogues out of 300 (21-26%), NMD rates Popularity higher than Uniform, disagreeing with the other measures.

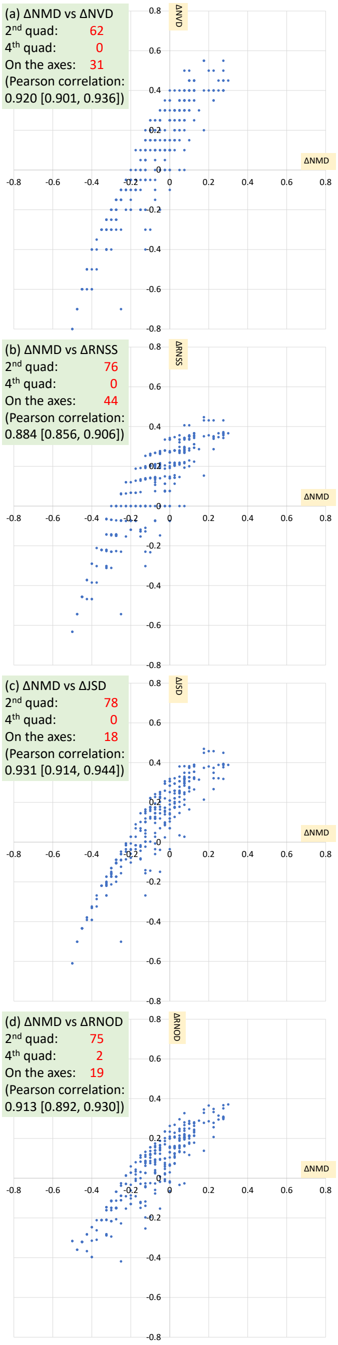

Figure 1: Scatterplot of per-dialogue score deltas (A_BL-popularity − A_BL-uniform from DialEval-1 DQ-A): NMD vs NVD/RNSS/JSD/RNOD. Number of instances in the 2nd and 4th quadrants, as well as Pearson correlations (with 95%CIs, $n = 300$), are also shown.

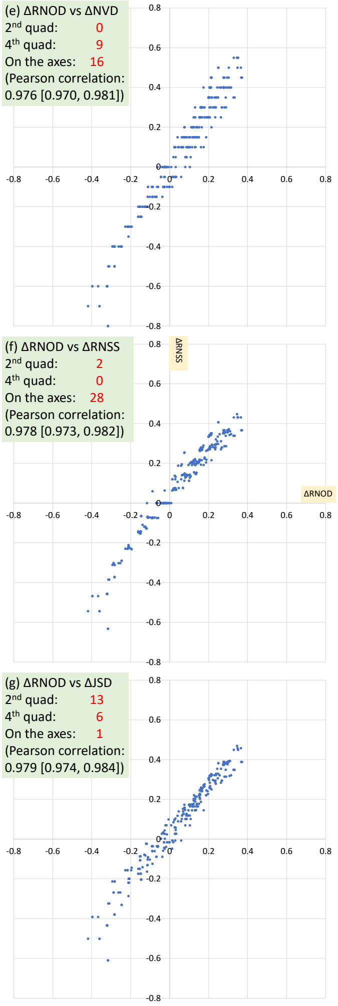

Figure 2: Scatterplot of per-dialogue score deltas (A_BL-popularity − A_BL-uniform from DialEval-1 DQ-A): RNOD vs NVD/RNSS/JSD. Number of instances in the 2nd and 4th quadrants, as well as Pearson correlations (with 95%CIs, $n = 300$), are also shown.

Table 7
Per-dialogue wins and losses between the Popularity baselines and the Uniform baselines. The higher number in each condition is shown in **bold**.

	NMD		RSNOD		RNOD		NVD		RNSS		JSD	
	uni	pop	uni	pop	uni	pop	uni	pop	uni	pop	uni	pop
STC-3 DQ-A	175	**197**	**224**	166	**265**	125	**274**	116	**297**	93	**301**	89
STC-3 DQ-E	**257**	127	**294**	96	**331**	58	**328**	62	**340**	50	**347**	43
STC-3 DQ-S	109	**268**	177	**213**	**229**	161	**224**	166	**257**	133	**255**	135
DialEval-1 DQ-A	97	**185**	**164**	136	**185**	114	**172**	113	**186**	86	**192**	108
DialEval-1 DQ-E	**128**	155	**174**	124	**213**	85	**198**	77	**226**	42	**229**	71
DialEval-1 DQ-S	79	**209**	133	**167**	**177**	123	**157**	120	**186**	90	**191**	109
TOTAL	845	**1141**	**1166**	902	**1400**	666	**1353**	654	**1492**	494	**1515**	555

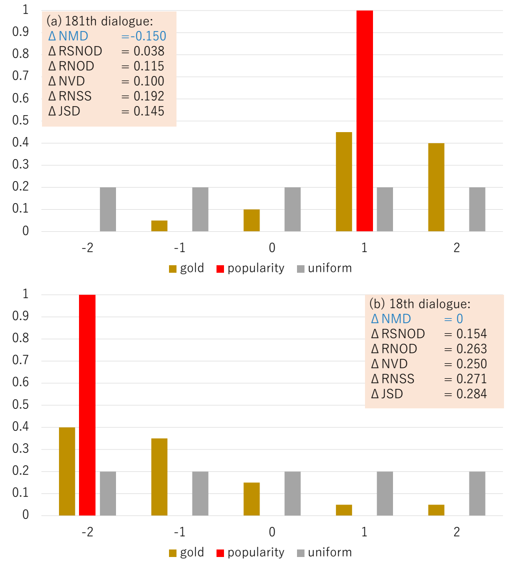

Figure 3: Top two dialogues from DialEval-1 DQ-A when ranked by $d = \Delta$NMD $-$ ΔRNOD: 181th dialogue ($d = -0.265$) and 18th dialogue ($d = -0.263$).

Whether a measure prefers Popularity or Uniform depends on what the gold distribution for each dialogue looks like. To closely examine situations where NMD favours Popularity while disagreeing with the other measures, we shall discuss two actual dialogues from the DialEval-1 DQ-A data below, which were selected as follows. First, because we are primarily interested in how and why NMD and RNOD behave differently, we ranked the 300 dialogues by how the ΔNMD and ΔRNOD values differ, that is, $d = \Delta\text{NMD} - \Delta\text{RNOD}$.

Figure 3 shows the gold, Popularity, and Uniform distributions for the top two dialogues in terms of d. In Figure 3(a) (181th dialogue, $d = -0.150-0.115 = -0.265$), it can be observed that Class 1 has the highest gold probability, and therefore that Popularity sets the probability of Class 1 to be 1. This is how Popularity "cheats." As shown in the pink box, all measures except NMD have positive ΔM's; that is, they say that Popularity underperforms Uniform. In contrast, NMD prefers Popularity. (Recall that for quantification measure scores, smaller means better.) In Figure 3(b) (18th dialogue, $d = 0 - 0.263 = -0.263$), Class -2 has the highest gold probability. For this dialogue, NMD says that Popularity and Uniform are equally effective, while all other measures prefer Uniform. It is clear from these examples that it is difficult to say whether one measure is "correct" or not; we can only say that NMD tends to prefer Popularity over Uniform compared to the other measures.

Using the DialEval-1 DQ-A data set, we have so far discussed how NMD tends to favour Popularity over Uniform. To generalise this observation, Table 7 shows how often each measure prefers one of the two baselines, for each of the six NTCIR data sets that contain these baselines. For example, NMD prefers Uniform for 175 dialogues and prefers Popularity for 197 dialogues from the DTC-3 DQ-A data set. (For the remaining $390 - 175 - 197 = 18$ dialogues, the two baselines are tied.) It can be observed from the TOTAL row that while RNOD, NVD, RNSS, and JSD prefer Uniform far more often (where the probability that Popularity is preferred is far below 50%), NMD prefers Popularity more often than it prefers Uniform. As for RSNOD, it does prefer Uniform more often just like RNOD and others, but the tendency is less clear; again, its property lies somewhere between NMD and RNOD.

6. Conclusions

The present study re-examined the OQ measures (NMD and R(S)NOD) along with nominal quantification measures (NVD, RNSS, and JSD) using SemEval and NTCIR data sets, using statistical significance test results. Our main findings are as follows.

- According to our pooled discriminative power results (Table 3), the most discriminative measures are RNOD and NVD (but recall that NVD is a nominal quantification measure and is not appropriate for OQ).
- The sets of statistically significant differences found by NMD are generally *not* subsets of those found by other, more discriminative measures like RNOD.
- NMD sometimes contradicts with RNOD and the nominal quantification measures in statistical terms, by preferring a Popularity baseline over a Uniform baseline.

The tendency of NMD to rate Popularity higher than Uniform is generally intuitive, since the former "cheats" by accessing the gold data while the latter is the laziest approach possible. However, it is difficult to say whether NMD is more appropriate than RNOD, as the preference between Popularity and Uniform should depend on what the gold distribution looks like (e.g., Is it almost flat?). On the other hand, the strengths of RNOD are that it is statistically stable, as demonstrated in terms of system ranking consistency [1] and pooled discriminative power. Based on these arguments, we recommend using both RNOD and NMD for evaluating OQ systems, to examine them from multiple angles.

Our future work includes exploring variants of RNOD. More specifically, while Eq. 2 relies on $\delta_{ij} = |i - j|$ and therefore assumes equidistance, an alternative δ_{ij} that is free from this assumption could be considered. Inspired by the distance function used in Krippendorff's alpha for ordinal classes [21, 22], one possibility is:

$$\delta_{ij} = \left(\sum_{k=\min(i,j)}^{\max(i,j)} p^*_k \right) - \frac{p^*_i + p^*_j}{2} \,. \tag{11}$$

That is, we could utilise the gold propabilities that lie between Classes i and j to define the distance. This can also be combined with the RNADW measure that we defined in Section 3.

Acknowledgments

We thank the reviewers of the LQ 2021 workshop for their feedback on the initial version of this paper, and the organisers of the workshop for giving us the opportunity to publish our work.

A. Proof That RNOD equals NMD when $|C| = 2$.

Note that $cp_1 = p_1$ and $cp^*_1 = p^*_1$ in general. Furthermore, when $|C| = 2$, note that $cp_2 = cp^*_2 = 1$. Hence, from Eq. 1,

$$\begin{aligned} NMD(p, p^*) &= |cp_1 - cp^*_1| + |cp_2 - cp^*_2| \\ &= |p_1 - p^*_1| + 0 = |p_1 - p^*_1| \,. \end{aligned} \tag{12}$$

On the other hand, note that when $|C| = 2$, $(p_2 - p^*_2)^2 = (1 - p_1 - 1 + p^*_2)^2 = (p_1 - p^*_1)^2$. To compute RNOD, the following three cases need to be considered. **Case 1** when $p^*_1 > 0$ *and* $p^*_2 > 0$: from Eq. 3, $OD(p \mid\mid p^*) = (DW_1 + DW_2)/2 = ((p_2 - p^*_2)^2 + (p_1 - p^*_1)^2)/2 = 2(p_1 - p^*_1)^2/2 = (p_1 - p^*_1)^2$. Hence from Eq. 5,

$$RNOD(p \mid\mid p^*) = \sqrt{OD(p \mid\mid p^*)} = |p_1 - p^*_1| \,. \tag{13}$$

Case 2 when $p^*_1 = 1$ and $p^*_2 = 0$: from Eq. 3, $OD(p \mid\mid p^*) = DW_1 = (p_2 - p^*_2)^2 = (p_1 - p^*_1)^2$. Therefore, Eq 14 holds for this case as well. **Case 2** when $p^*_1 = 0$ and $p^*_2 = 1$: from Eq. 3, $OD(p \mid\mid p^*) = DW_1 = (p_1 - p^*_1)^2$ and Eq 14 holds for this case as well. In summary, $NMD(p, p^*) = RNOD(p \mid\mid p^*)$.

Finally, following similar steps as above, we can also obtain:

$$RNOD(p^* \mid\mid p) = \sqrt{OD(p^* \mid\mid p)} = |p_1 - p^*_1| \,. \tag{14}$$

In summary, $NMD(p, p^*) = RNOD(p \mid\mid p^*) = RNOD(p^* \mid\mid p)$ when $|C| = 2$. Q.E.D.

References

[1] T. Sakai, Evaluating evaluation measures for ordinal classification and ordinal quantification, in: Proceedings of ACL-IJCNLP 2021, 2021, pp. 2759–2769. URL: https://aclanthology.org/2021.acl-long.214.pdf.

[2] A. Esuli, F. Sebastiani, Sentiment quantification, IEEE Intelligent Systems 25 (2010) 72–75.

[3] W. Gao, F. Sebastiani, From classification to quantification in tweet sentiment analysis, Social Network Analysis and Mining 6 (2016) 1–22.

[4] F. Sebastiani, Evaluation measures for quantification: an axiomatic approach, Information Retrieval Journal 23 (2020) 255–288.

[5] P. Nakov, A. Ritter, S. Rosenthal, V. Stoyanov, F. Sebastiani, SemEval-2016 task 4: Sentiment analysis in Twitter, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, Association for Computational Linguistics, San Diego, California, 2016. URL: https://www.aclweb.org/anthology/S16-1001.pdf.

[6] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval '17, Association for Computational Linguistics, Vancouver, Canada, 2017. URL: https://www.aclweb.org/anthology/S17-2088.pdf.

[7] R. Higashinaka, K. Funakoshi, M. Inaba, Y. Tsunomori, T. Takahashi, N. Kaji, Overview of Dialogue Breakdown Detection Challenge 3, in: Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop, 2017. URL: http://workshop.colips.org/dstc6/papers/track3_overview_higashinaka.pdf.

[8] Z. Zeng, S. Kato, T. Sakai, Overview of the NTCIR-14 short text conversation task: Dialogue quality and nugget detection subtasks, in: Proceedings of NTCIR-14, 2019, pp. 289–315. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-STC-ZengZ.pdf.

[9] Z. Zeng, S. Kato, T. Sakai, I. Kang, Overview of the NTCIR-15 dialogue evaluation task (DialEval-1), in: Proceedings of NTCIR-15, 2020, pp. 13–34. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/01-NTCIR15-OV-DIALEVAL-ZengZ.pdf.

[10] M. Werman, S. Peleg, A. Rosenfeld, A distance metric for multidimensional histograms, Computer Vision, Graphics, and Image Processing 32 (1985) 328–336.

[11] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision 40 (2000) 99–121.

[12] E. Levina, P. Bickel, The earth mover's distance is the mallows distance: Some insights from statistics, in: Proceedings of ICCV 2001, 2001, pp. 251–256.

[13] T. Sakai, Towards automatic evaluation of multi-turn dialogues: A task design that leverages inherently subjective annotations, in: Proceedings of EVIA 2017, 2017, pp. 24–30. URL: http://ceur-ws.org/Vol-2008/paper_4.pdf.

[14] T. Sakai, Comparing two binned probability distributions for information access evaluation, in: Proceedings of ACM SIGIR 2018, 2018, pp. 1073–1076. URL: https://dl.acm.org/doi/pdf/10.1145/3209978.3210073.

[15] T. Sakai, On the instability of diminishing return IR measures, in: Proceedings of ECIR 2021 Part I (LNCS 12656), 2021, pp. 572–586.

[16] T. Sakai, Evaluating evaluation metrics based on the bootstrap, in: Proceedings of ACM SIGIR 2006, 2006, pp. 525–532.

[17] T. Sakai, Alternatives to bpref, in: Proceedings of ACM SIGIR 2007, 2007, pp. 71–78.

[18] Y. Tsunomori, R. Higashinaka, T. Takahashi, M. Inaba, Selection of evaluation metrics for dialogue breakdown detection in dialogue breakdown detection challenge 3 (in Japanese), Transactions of the Japanese Society for Artificial Intelligence 35 (2020). URL: https://www.jstage.jst.go.jp/article/tjsai/35/1/35_DSI-G/_pdf/-char/ja.

[19] T. Sakai, Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power, Springer, 2018.

[20] T. Sakai, Metrics, statistics, tests, in: PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173), Springer, 2014, pp. 116–163.

[21] K. Krippendorff, Content Analysis: An Introduction to Its Methodology (Fourth Edition), SAGE Publications, 2018.

[22] T. Sakai, How to run an evaluation task, in: Information Retrieval Evaluation in a Changing World, Springer, 2019, pp. 71–102.