Analysis of Measurement Precision Experiment with Categorical Variables

Tomomichi Suzuki, Jun-ichi Takeshita, Mayu Ogawa, Xiao-Nan Lu and Yoshikazu Ojima

Abstract Evaluating performance of a measurement method is essential in metrology. Concepts of repeatability and reproducibility are introduced in ISO5725-1 (1994) including how to run and analyse experiments (usually collaborative studies) to obtain these precision measures. ISO5725-2 (1994) describe precision evaluation in quantitative measurements but not in qualitative measurements. Some methods have been proposed for qualitative measurements cases such as Wilrich (2010), de Mast & van Wieringen (2010), Bashkansky, Gadrich & Kuselman (2012). Item response theory (Muraki, 1992) is another methodology that can be used to analyse qualitative data. Utilizing these methods, analysis of measurement precision experiment is investigated when the measurements involve categorical variables.

The data analysed are from the precision experiment of intratracheal administration testing (AIST, 2018) whose objectives were to study the precision of the standardized test method for evaluating the pulmonary toxicity of nanomaterials. In such experiments, dose-response relationship also need to be considered which

Tomomichi Suzuki
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: szk@rs.tus.ac.jp

Jun-ichi Takeshita
Research Institute of Science for Safety and Sustainability, National Institute of Advanced Industrial Science and Technology (AIST), 16-1 Onogawa, Tsukuba, Ibaraki, 305-8569, Japan, e-mail: jun-takeshita@aist.go.jp

Mayu Ogawa
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: 7417605@ed.tus.ac.jp

Xiao-Nan Lu
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: lu@rs.tus.ac.jp

Yoshikazu Ojima
Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, e-mail: ojima@rs.tus.ac.jp

makes the situation more complicated. We discuss how these data should be analysed using actual data.

1 Introduction

Evaluating performance of a measurement method is essential in metrology. Concepts of repeatability and reproducibility are introduced in ISO5725-1 (1994) including how to run and analyse experiments (usually collaborative studies) to obtain these precision measures. ISO5725-2 (1994) describes precision evaluation in quantitative measurements but not in qualitative measurements. Some methods have been proposed for qualitative measurements cases such as Wilrich (2010), de Mast & van Wieringen (2010), Bashkansky, Gadrich & Kuselman (2012). Item response theory (Muraki, 1992) is another methodology that can be used to analyse qualitative data. Utilizing these methods, analysis of measurement precision experiment is investigated when the measurements involve categorical variables.

The data analysed are from the precision experiment of intratracheal administration testing (AIST, 2018) whose objectives were to study the precision of the standardized test method for evaluating the pulmonary toxicity of nanomaterials. In such experiments, dose-response relationship also need to be considered which makes the situation more complicated. The objective of this paper is to discuss how these data should be analysed using actual data.

Section 2 explains the data used in this paper. In Section 3, the methods for analyzing qualitative data are introduced. In Section 4, the results of the analyses are described. Section 5 gives the summary.

2 Data

The data used in this paper are from the precision experiment of intratracheal administration testing (AIST, 2018). The main objective of the experiment was to study the precision of the standardized test method for evaluating the pulmonary toxicity of nanomaterials. The overview of the experiment is as follows:

a) There were three nanomaterials used in the study.
b) For each nanomaterial, four levels of dose were designed (none, low, middle, high).
c) The experiment was carried out by five laboratories.
d) The number of replicates was five.
e) The replications were carried out by using rats.
f) The same test method is used for all the laboratories.
g) The equipments used in each laboratory may be different.
h) Each rat went through pathological examination.
i) There are 19 characteristics to be examined by experts.

j) The result of the examination reveals the inflammation grade of response.
k) The grade of response is one of five grades ($-$, $+-$, $+$, $++$, $+++$).

Therefore, we obtain categorical data (number of category is five) of 19 characteristics for each of five rats, four level of doses, three nanomaterials in each laboratory.

3 Methods

3.1 ISO 5725

In ISO 5725, accuracy of a measurement result, measurement method, or measurement system is a general term that involves trueness and precision. Trueness, the closeness of agreement between the average value obtained from a large series of measurement results and an accepted reference value, is usually expressed in terms of bias, which is the difference between expectation of the measurement results and accepted reference value. Precision, the closeness of agreement between independent measurement results obtained under stipulated conditions, is usually expressed in terms of standard deviations of the measurement results.

Generally, when the accuracy of a measurement method is to be repeated, it is known that two measures of accuracy, named repeatability and reproducibility, are required. Repeatability is measurement results under repeatability conditions, where the independent measurement results are obtained using the same method on the identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. Reproducibility is measurement results under reproducibility conditions, where the measurement results are obtained using the same method on identical test items in different laboratories with different operators using different equipment.

In the ISO 5725 series, the basic model for a measurement result y is given by

$$y = m + B + e \tag{1}$$

to estimate accuracy of measurement method. m is general mean (expectation), B is laboratory component of variation (under repeatability conditions), e is random error (under repeatability conditions). The expectation of B is assumed to be 0, and the variance of B, which is the between-laboratory variance, is shown by σ_L^2. The expectation of e is assumed to be 0, and the variance of e, which is the within-laboratory variance, is assumed to be equal in all laboratories and is denoted as repeatability variance σ_r^2. Reproducibility variance σ_R^2 can be expressed as the sum of between-laboratory variance and repeatability variance, that is

$$\sigma_R^2 = \sigma_L^2 + \sigma_r^2. \tag{2}$$

3.2 Ordinal Analysis of Variance (ORDANOVA)

ORDANOVA (Bashkansky, Gadrich & Kuselman, 2012) is a method to investigate the difference of measurement results among laboratories when the results are measured as ordered categorical data. In ORDANOVA, the null hypothesis is defined as "there are no difference in measurement results among all the laboratories" and the alternative hypothesis is defined as "there are differences in measurement results among some of the laboratories".

The within-laboratory variation $\widehat{h}^2_{m(W)}$ is given as Eq. (3).

$$\widehat{h}^2_{m(W)} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \widehat{F}_{km}(1-\widehat{F}_{km}), \tag{3}$$

where $\widehat{F}_{km}$ denotes the cumulative frequency of data with laboratory $m = 1,2,\ldots,M$ and category $k = 1,2,\ldots,K$. Measure of the between samples variation per kth category $\widehat{S}^2_{k(B)}$ is given by Eq. (4).

$$\widehat{S}^2_{k(B)} = \frac{1}{M} \sum_{m=1}^{M} \left(\widehat{F}_{km} - \widehat{F}_{k\cdot}\right)^2. \tag{4}$$

The total variation $\widehat{h}^2_{(T)}$ is given by Eq. (5).

$$\begin{aligned} \widehat{h}^2_{(T)} &= \widehat{h}^2_{(W)} + \widehat{S}^2_{(B)} \\ &= \frac{1}{M} \sum_{m=1}^{M} \widehat{h}^2_{m(W)} + \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \widehat{S}^2_{k(B)}. \end{aligned} \tag{5}$$

$\widehat{h}_{(W)}$, $\widehat{S}_{(B)}$ and $\widehat{h}_{(T)}$ are analogous to repeatability variance, reproducibility variance and total variance. The test statistic is

$$I = \frac{\widehat{S}^2_{(B)}/df_B}{\widehat{h}^2_{(T)}/df_T} \tag{6}$$

with degree of freedom $df_B = M-1$, $df_T = N-1$. The null hypothesis is rejected when

$$I > I_{cr} = \frac{\chi^2_{1-\alpha}}{M-1}, \tag{7}$$

where I_{cr} is the critical value and α is the significance level.

3.3 Attribute Agreement Analysis (AAA)

AAA (ISO-TR-14468, 2010) is a method to analyze agreement among nominal data. Fleiss' κ statistic is applied to investigate the agreement of measurements between and within laboratories. The estimate of κ is given by Eq. (8).

$$\widehat{\kappa} = \frac{\widehat{P}_o - \widehat{P}_e}{1 - \widehat{P}_e}, \tag{8}$$

where $\widehat{P}_o$ is the probability the measurement results actually matched and $\widehat{P}_e$ is the probability that the measurement results match by chance. κ takes the value between -1 and $+1$, and indicates more agreement if κ is nearer to $+1$.

To obtain agreement within laboratories, $\widehat{P}_o$ and $\widehat{P}_e$ expressed as Eq. (9) are used.

$$\begin{aligned} \widehat{P}_o &= \frac{1}{Nl(l-1)} \left(\sum_{i=1}^{N} \sum_{k=1}^{K} x_{ik}^2 - Nl \right), \\ \widehat{P}_e &= \sum_{k=1}^{K} p_k^2, \qquad p_k = \frac{1}{Nl} \sum_{i=1}^{N} x_{ik}. \end{aligned} \tag{9}$$

where x_{ik} denotes the frequency of item $i = 1,2,\ldots,N$ classified as category $k = 1,2,\ldots,K$, and l express the number of replications. To obtain agreement between laboratories, $\widehat{P}_o$ and $\widehat{P}_e$ expressed as Eq. (10) are used.

$$\widehat{P}_o = \frac{1}{NMl(Ml-1)} \left(\sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{M} x_{ijk}^2 - NMl \right), \tag{10}$$

where x_{ijk} denotes the frequency of item $i = 1,2,\ldots,N$ classified as category $k = 1,2,\ldots,K$ for laboratory $j = 1,2,\ldots,M$.

3.4 Item Response Theory (IRT)

IRT (AIST, 2018) has been developed in the field of education and psychology. It is widely used in tests and examinations. It enables the user to estimate both the ability of the examinees and the difficulty of the questions.

Many models are developed in IRT in order to accommodate various situations. Generalized Partial Credit Model (GPCM) is a model applied for ordinal polytomous data, which the partial point can be possible for each test question. GPCM is expressed as Eq. (11)

$$P_{jh}(\theta) = \frac{\exp\left\{a_j \sum_{m=0}^{h} (\theta - b_{jm})\right\}}{\sum_{h=0}^{H} \exp\left\{a_j \sum_{m=0}^{h} (\theta - b_{jm})\right\}}, \tag{11}$$

where $P_{jh}(\theta)$ denotes the probability of examinee with ability θ obtaining partial points h for test question (called item in IRT). The plots of the probabilities with regard to ability are called Item Characteristic Curve (ICC) and an example is shown in Fig. 1. In Fig. 1, b_{jm} are the ability values of the intersections (thresholds) of adjacent points. Larger b_{jm} means that a questions is more difficult. In Fig. 1, a_j express the slope of the tangents at the intersections. Larger a_j means that a question discriminates examinees better. When IRT is applied to precision experiments, the

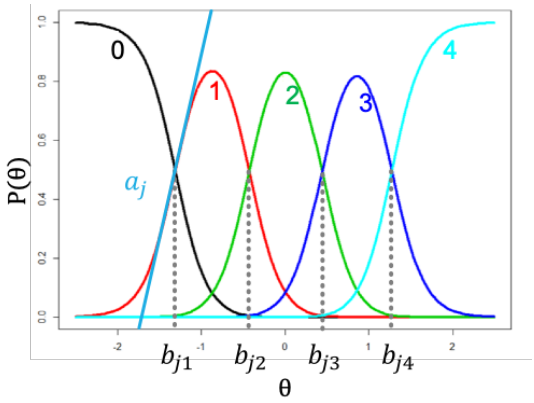


Fig. 1: Item Characteristic Curve for item j

model can be expressed as Eq. (12) (AIST, 2018).

$$q_j(h|x) = \frac{\exp\left\{\alpha_j \sum_{m=0}^{h}(x-\delta_{jm})\right\}}{\sum_{h=0}^{H}\exp\left\{\alpha_j \sum_{m=0}^{h}(x-\delta_{jm})\right\}}, \tag{12}$$

Here, $q_j(h|x)$ denotes the probability that a laboratory j gives category h when the toxity of the nanomaterial is x. α_j are the discrimination parameters and δ_{jm} are the threshold parameters. Laboratories in precision experiments are regarded as test questions in IRT, and the toxity of the nanomaterial in precision experiments are regarded as ability of examinees in IRT.

Precision measures for the variation within a laboratory and between laboratories are given by Eq. (13) and Eq. (14) respectively (AIST, 2018).

$$\pi_j^w(h) = P\left(Y_{ij} = h \mid \delta_{j,h} < X_i < \delta_{j,h+1}\right), \tag{13}$$

$$\pi_{j_1,j_2}^b = \sum_{h=0}^{H} P\left(\delta_{j_1,h-1} < X < \delta_{j_1,h} \wedge \delta_{j_2,h-1} < X < \delta_{j_2,h}\right), \tag{14}$$

Repeatability can be calculated by taking the average of $\pi_j^w(h)$ and reproducibility can be calculated by taking the average of π_{j_1,j_2}^b.

4 Results and discussion

4.1 Graphical Presentation

The data structure of the toxicity experiment was complicated. We have started with drawing bubble charts. To clarify the dose-response relationship, bubble charts are applied. An example of the bubble charts are shown in Fig. 2. Size of a bubble correspond to frequency, which is the number of rats. Those graphs were drawn for all the possible cases, that is 285 (3 nanomaterials $\times$ 19 characteristics $\times$ 5 laboratories).

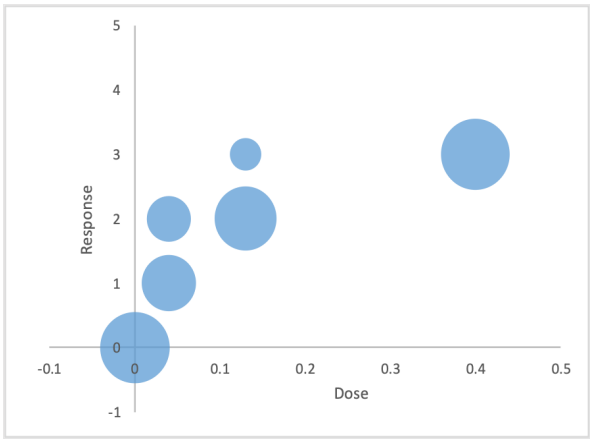


Fig. 2: Bubble Chart (Nanomaterial C, Characteristic No.1, Laboratory No.1)

4.2 Estimation of Precision Measures

Repeatability and reproducibility measures are calculated for each of the combinations of nanomaterials, doses and characteristics using quantitative ISO 5725 method. Significance tests were carried out to check if the existence of between-laboratory variance. An example of the summarized results (for nanomaterial A) is

shown in Table 1. More significant results are obtained for higher doses. If we focus on F-values, they are larger for characteristics Nos. 8, 9, 11 and 18.

Table 1: Results of ISO 5725 (Nanomaterial A)

Char	None	Low	Medium	High
1	0.0	8.6*	4.1	4.8
2	0.0	3.4	5.7	10.9*
6	0.0	2.4	9.6*	6.0
7	0.0	6.0*	5.7	4.3
8	2.0	5.1	22.0*	17.6*
9	1.0	2.1	19.6*	12.9*
10	4.0	3.4	3.8	4.8
11	0.0	18.6*	5.7	14.7*
16	0.0	1.1	0.0	2.7
18	0.0	1.0	7.5*	23.5*
19	0.0	0.0	2.7	1.0

*: statistically significant

Repeatability and reproducibility measures are calculated for each of the combinations of nanomaterials, doses and characteristics using ORDANOVA. The summarized results for nanomaterial A is shown in Table 2. The repeatability measures are larger for characteristics Nos. 10 and 11. The reproducibility measures are larger with higher doses, especially for the characteristic No. 2. Repeatability and repro-

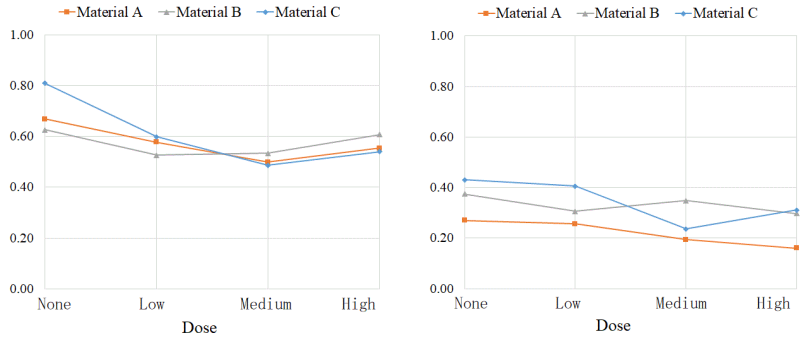
Table 2: Results of ORDANOVA (material A)

	None		Low		Medium		High	
Char.	r	R	r	R	r	R	r	R
1	0.00	0.00	0.08	0.25	0.08	0.27	0.00	0.40
2	0.00	0.00	0.08	0.25	0.11	0.52	0.00	0.72
6	0.13	0.28	0.06	0.28	0.13	0.29	0.10	0.39
7	0.00	0.00	0.10	0.30	0.22	0.49	0.16	0.51
8	0.05	0.07	0.18	0.32	0.18	0.60	0.18	0.48
9	0.10	0.20	0.08	0.26	0.19	0.42	0.16	0.49
10	0.13	0.36	0.21	0.23	0.26	0.31	0.26	0.36
11	0.05	0.07	0.16	0.21	0.24	0.45	0.26	0.50
16	0.06	0.08	0.06	0.08	0.03	0.04	0.16	0.22
18	0.05	0.11	0.03	0.20	0.08	0.38	0.10	0.48
19	0.00	0.00	0.06	0.08	0.21	0.38	0.19	0.42

r: repeatability, R: reproducibility

ducibility measures are calculated for each of the combinations of nanomaterials, doses and characteristics using nominal AAA. The summarized results for nanoma-

terial A is shown in Fig. 3. The variation becomes larger with higher doses. The



(a) Repeatability (b) Reproducibility

Fig. 3: Results of AAA (material A)

results above are aggregated as Table 3. The results among ISO 5725, ORDANOVA and AAA were compared. From Table 3, it became clear that the peak of the precision measures are the same among all the methods. The variations was smallest with no dose in all the cases. Repeatability and reproducibility measures are cal-

Table 3: Aggregated Results

		Repeatability			Reproducibility		
		5725	ORD.	AAA	5725	ORD.	AAA
Material A	None	0.08	0.05	0.67	0.14	0.11	0.27
	Low	0.15	0.10	0.58	0.32	0.22	0.26
	Medium	0.28	0.16	0.50	0.64	0.38	0.19
	High	0.21	0.14	0.55	0.67	0.45	0.16
Material B	None	0.09	0.05	0.63	0.08	0.08	0.37
	Low	0.33	0.16	0.53	0.62	0.32	0.31
	Medium	0.25	0.15	0.53	0.61	0.32	0.35
	High	0.30	0.15	0.61	0.73	0.36	0.30
Material C	None	0.03	0.02	0.81	0.04	0.05	0.43
	Low	0.16	0.10	0.60	0.26	0.18	0.41
	Medium	0.27	0.16	0.49	0.69	0.35	0.24
	High	0.23	0.15	0.54	0.67	0.34	0.31

culated for each nanomaterial using IRT. ICCs (Item Characteristic Curves) were drawn for all the laboratories for each nanomaterial. Figure 4 shows ICC of laboratory No. 1 for nanomaterial A. The vertical axis expresses the probability of the measurement result being a specific category h, which is denoted by $q_j(h|x)$. The horizontal axis expresses the true toxicity which is denoted by x. From Fig. 4, the

probability of the measurement result being category 1 is more than ninety five percent when $x = 0$. We also can read that when $x = 2$, the probability of category 3 is about 40% and the probability of category 4 is about 60%. Table 4 shows the results

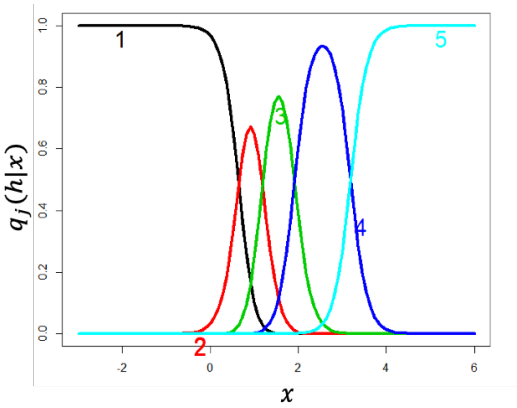


Fig. 4: Item Category Curve (material A, Lab No.1)

of the estimated parameters for nanomaterial A. The order of the estimates of the discrimination parameter α_j were Labs No.1, No.2, No.3, No.4, No.5 in descending order, which means appropriate classification were also performed in this order from better to worse. The estimates of the threshold parameters were generally similar among laboratories except for Lab No.3, which suggest there exist some kind of bias. Repeatability and reproducibility using Eqs. (13) and (14) are calculated

Table 4: Estimate of discrimination parameters and threshold parameters

	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
δ_{j1}	0.65	0.74	1.01	0.68	0.75
δ_{j2}	1.18	1.29	1.68	0.86	1.00
δ_{j3}	1.91	1.80	3.04	1.70	1.33
δ_{j4}	3.19	3.24	3.00	3.38	2.19
α_j	5.25	4.22	3.10	3.06	1.61

using the threshold parameters shown in Table 4. Table 5 is the repeatability and reproducibility measures for all the nanomaterials. The repeatability was highest with nanomaterial C and the reproducibility was highest with nanomaterial B. Comparing repeatability and reproducibility, reproducibility was larger with nanomaterials A and B, while repeatability was larger with nanomaterial C.

Table 5: Precisions of the measurement method

	Material A	Material B	Material C
Repeatability	0.808	0.817	0.829
Reproducibility	0.847	0.854	0.777

4.3 Estimation of Toxicity

Toxicity of the nanomaterials are estimated using IRT. Fig. 5 is box-whisker plots of the estimated toxicity with material A. From Fig. 5, we can see the effect of doses. Although the distribution of the toxicity among the levels of doses overlap considerably, the effect of doses are apparent as a whole, that is, the average of the toxicity becomes larger as dose become larger.

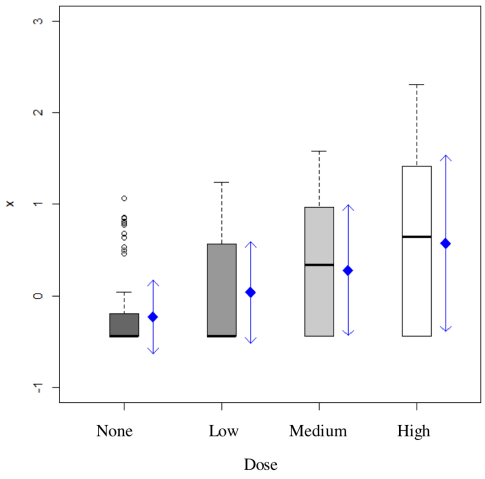


Fig. 5: Box-whisker plot of estimated toxicity (material A)

5 Summary

The precision measures of ordinal categorical data were estimated for actual intratracheal administration testing experiment data. The dose-response relationships were also investigated. It was not able to perform an analysis that takes account of

all the factors as there were many factors in the experiment. Therefore, analyses were performed by considering appropriate factors for each analysis.

Precision measures for characteristics were well clarified by ISO 5725 and ORDANOVA. The characteristics which have large variations were identified. Precision measures for each dose were estimated by ISO 5725, ORDANOVA and AAA. It became clear that the variation becomes larger as the dose becomes larger. The condition that gave the maximum and the minimum values were the same. The precision measures for each nanomaterials are estimated using IRT. The repeatability were larger with material B and the reproducibility were larger with material C. The features of each laboratory could be observed using Item Characteristic Curve of IRT. The dose-response relationships were examined using the estimated toxicity using IRT. The relationships were investigated for each characteristics for each nanomaterial. Due to the inherent variation in the data, it was not possible to obtain the precise dose-response relationship. Nevertheless, the existence of dose-response relationship could be verified using rank correlation coefficient.

References

International Organization for Standardization (1994). ISO 5725-1 Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions. Geneva: ISO.

International Organization for Standardization (1994). ISO 5725-2 Accuracy (trueness and precision) of measurement methods and results – Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. Geneva: ISO.

P.-Th. Wilrich, (2010). "The determination of precision of qualitative measurement methods by interlaboratory experiments". Accreditation and Quality Assurance **15**, 439–444.

J. de Mast and W. van Wieringen, (2010). "Modeling and Evaluating Repeatability and Reproducibility of Ordinal Classifications". Technometrics **52**(1), 94–106.

E. Bashkansky, T. Gadrich and I. Kuselman, (2012). "Interlaboratory comparison of test results of an ordinal or nominal binary: analysis of variation". Accreditation and Quality Assurance **17**, 239–444.

E. Muraki, (1992). "A Generalized Partial Credit Model: Application of an EM Algorithm". Applied Psychological Measurement **16**(2), 159–176.

The National Institute of Advanced Industrial Science and Technology (AIST), (2018). "Annual Report on a Project 'Survey on standardization of intratracheal administration study for nanomaterials and related issues' ". In Japanese, accessed 2019-05-31: http://www.meti.go.jp/meti_lib/report/H29FY/000102.pdf.

International Organization for Standardization (ISO), (2010). ISO/TR 14468 Selected illustrations of attribute agreement analysis. Geneva: ISO.