XIX IMEKO World Congress
Fundamental and Applied Metrology
September 6−11, 2009, Lisbon, Portugal

ATTEMPTS TO DIMINISH UNCERTAINTY IN QUALITY EVALUATION OF COMPRESSED VIDEO BY HUMAN AUDIENCE

Anna Ostaszewska [1], *Sabina Żebrowska-Łucyk* [2]

[1] Warsaw University of Technology, Warsaw, Poland, a.ostaszewska@mchtr.pw.edu.pl
[2] Warsaw University of Technology, Warsaw, Poland, s.zebrowska@mchtr.pw.edu.pl

Abstract – The paper concerns one of subjective continuous quality evaluation method (SSCQE) [1, 2], which is used for measuring the human perception of compression errors in video. The problem with subjective quality assessment is a big measurement uncertainty that makes drawing conclusions on quality of compared materials with required significance level difficult. The paper presents a new method of measurement data processing, which enables for decreasing the unwanted influence of human factors and diminishing the standard deviation of the measurement results by a half. The Mandel's *h* and *k* statistics are used, likewise in inter-laboratory comparisons programs.

Keywords: SSCQE, Mandel's statistics, data filtration in subjective measurements

1. INTRODUCTION

Dynamic expansion of the Internet and wide variety of multimedia services involves a strong demand on effective lossy compression algorithms. Lossy compression process makes the output files smaller than the original, due to omitting some part of information, particularly redundant or of little importance for viewers. However, very often the differences between the source video and the compressed one result in visible distortions, called compression errors. The amount of impairments depends not only on algorithms and compression parameters, but also on the video content. The increasing use of compressed video calls for monitoring and assessment of the picture quality. The progress in the area of compression techniques is conditional upon the development of methods of quality evaluation.

Picture quality assessment depends on individual human being perception, so the appropriate way to obtain valuable results is to conduct evaluation with a panel of observers. Working with a human audience (usually non-experts), which assesses the quality perceived during a test session when the variously compressed video sequences are displayed, is the idea of subjective quality evaluation methods. The scores given by individual observers are processed to calculate the mean opinion scores (MOS).

Among methods for subjective quality evaluation the most interesting are continuous methods, which enable for acquisition and recording the series of subjects' opinions in time. Recorded signals yield the information not only on the global video quality itself, but also its temporary variations. Therefore they are both the source of knowledge on Human Visual System (HVS) and the most reliable tool to assess the performance of different compression algorithms. Hence for both scientific and practical reasons they ought to be developed and validated.

Despite the fact that numerous laboratories in the world use continuous methods [3 - 12], they are poorly examined from the metrological point of view. There are very few concepts how to estimate and lessen influence of undesirable random factors on measurement signals [2] and how to appraise and "calibrate" a single-person response.

2. SSCQE METHOD - PROPERTIES AND DRAWBACKS

The leading subjective continuous method is the Single Stimulus Continuous Quality Evaluation (SSCQE) recommended by International Telecommunication Union (ITU) [1, 2]. It is a non-reference way of quality assessment, which means that the audience watches the compressed video only, without the source video simultaneously given. This way is close to home conditions in which the video is to be watched. The panel of viewers should be both large (at least 15 viewers, according to ITU) and homogeneous enough for determination of statistically reliable scores.

The SSCQE method considers long-duration sequences (3 to 30 min). To assess video quality, each viewer operates with a slider device, with a 0-100 scale attached. The slider is connected to a PC and its position is sampled twice a second.

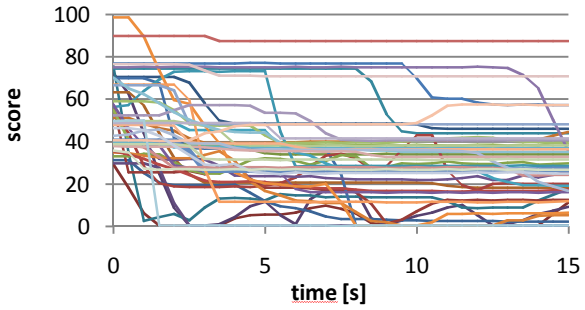

Fig. 1. A section of measurement signals from a group of observers, SSCQE method.

ISBN 978-963-88410-0-1

Although all viewers watch the same test material, there are many disparities between their individual plots of scores given in time (Fig. 1). This to some degree is caused by the natural differences in individual characteristics of viewers: their visual perception and the ability to observe, sensitivity and tolerance to compression artifacts, their requirements and expectations, psychomotor skills like time of reaction and the interpretation of the semantic scale used [12, 13, 14, 15]. There are also various other phenomena that influence the individual shape of the score signal [4, 16]: unsymmetrical tracking, recency effect, memory limitations, negative peak effect, drift and many other, still waiting for discovery.

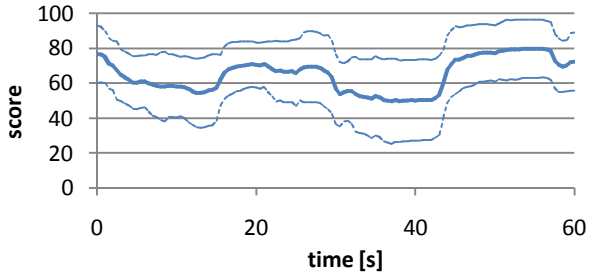

Fig. 2. A mean opinion score and two standard deviation intervals, SSCQE, 45 observers.

By the reason of strong differences between individual signals, confidence interval for mean opinion scores is very large (Fig. 2). This makes comparison of the quality of different parts of material difficult or even impossible. Therefore ITU-T recommends transforming original data to cumulative probability curves (Fig. 3).

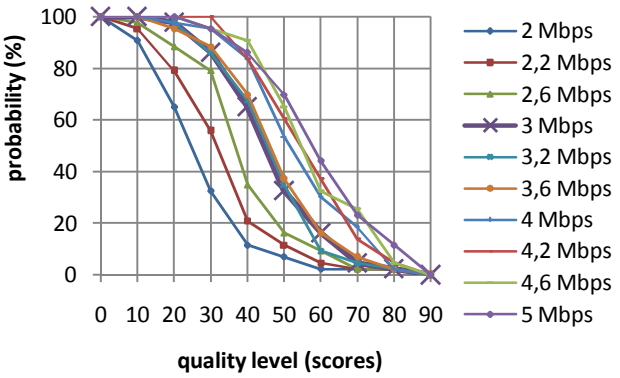

Fig. 3. Histogram of probability of the occurrence of quality level.

Unfortunately, such plots don't show temporal variability of scores given by the audience nor the confidence intervals. They are useful for comparison of different coders, but the whole information on HVS, necessary for enhancement of algorithms is lost.

There are two ways to reduce confidence interval for means. The first is to increase the number of observers but it is cost and time consuming. The better way is to detect and remove outliers. Recommendation [2] suggests critical approach to collected measurement data and it introduces a method for discarding incoherent data. However a number of experiments conducted by authors in Warsaw University of Technology revealed that the proposed method is not efficacious. In many cases it does not enable to exclude even obviously odd signals (e.g. completely incoherent with average signal).

Therefore we made an attempt to work out a new method of data processing that would diminish uncertainty intervals and thus would make subjective quality evaluation results more informative.

3. STATISTICAL FILTRATION OF MEASUREMENT SIGNALS

3.1. Assumptions

The raw data collected with the continuous method seems to be chaotic and it's rather impossible to explain fully the individual reaction to the material watched. The reasons for large standard deviations for mean values (apart from those which is impossible to take control of) are the following:

a) Some observers have an extremely weak ability of detecting and fast assessing the level of video didtortions.
b) Mean of scores and the range of scale used are individual for each observer and vary between subjects.
c) The dynamic of reaction to temporal variations of quality in time is an individual feature of each observer.
d) During a long test session, some periods with lower attention may occur, even in the case of attentive observers.

It was assumed that the new improved method of data processing should discard all signals which come from unreliable observers (a) and then lessen the influence of the next two phenomena (b, c), which are evidently natural and to discard the scores caused by the temporary lack of attention (d). Therefore it is rational to remove the whole measurement signals given by unreliable viewers and just small parts of accidentally distorted others signals.

Additionally the method should screen observers for their stability in assessing the quality: their scores should be coherent i.e. close in case of replicated evaluation of the same material. Thus it is necessary to modify slightly the measurement method itself: each observer assesses all video sequences twice. This enables for defining two kinds of data inconsistency: *inter* – the lack of consistency with the scores given by the same observer for the same test sequence and *intra* – the lack of consistency of scores given by the observer with the mean of scores given by the audience.

3.2. SSCQE experiment

For the purpose of collecting source data, needed to work out new filtration method, four 15-seconds sequences (Fig. 4) were coded in MPEG-2 with 10 levels of bitrate: from 2 to 5 Mbps, which is a typical range for this standard. The duration of test material was 10 minutes long and each observer (after a trial session) assessed the whole material twice. 45 male subjects, aged from 20 to 25, took part in the experiment.

Fig. 4. The screenshots of test sequences: ‘bbc3’, ‘mobl’, ‘cact’, ‘susi’.

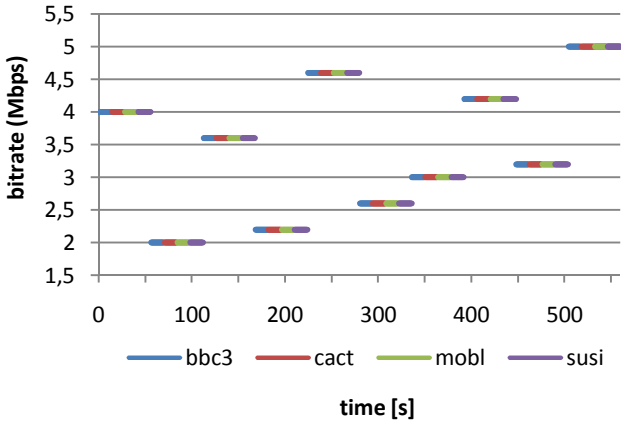

Fig. 5. The temporal layout of test material.

To be in accordance with ITU [2], the recommended filtration method was used, but as the result, no signal was qualified for rejection.

3.3. New filtration method

The new method applied for data filtration includes a series of operations that reduce the differences in the time of observers’ reaction, provides signal normalization and enables for objective rejection of parts of the signals or the whole of those, which are inter- or intra-inconsistent.

Due to the evident influence of the sequence content on the mean level of scores, each of four 15-seconds sequences were processed separately.

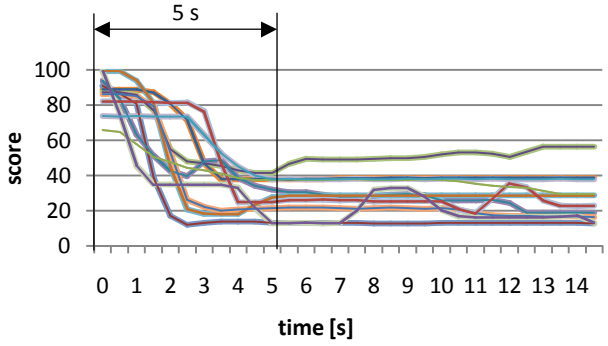

Fig. 6. Scores given in time by 10 observers for one 15-seconds sequence – first 5 seconds of voting is strongly influenced by the previous sequence.

In order to lessen the influence of dynamic of individual signals, the data filtration was based on the averages of scores calculated for each observer for a time of 10 s. The first 5 seconds of 15-seconds sequence were omitted, taking into account that viewers need approximately 5 seconds to adjust scores to new conditions (Fig. 6). From sets of individual observer’s scores, there were values of averaged scores computed, separately for each sequence, each bitrate and replication:

$$\bar{y}_{ijl.a} = \frac{1}{T - t'} \sum_{t=t'}^{T} y_{ijlta} \qquad (1)$$

where:
i – subscript for observer; $i = 1,\ldots, p$
j – subscript for level of coding; $j = 1,\ldots, 10$
l – subscript for replicate; $l = 1, \ldots, n$ ($n = 2$)
t – subscript for sample,
a – subscript for sequence,
T – number of all samples for the whole sequence,
$t’$ – number of samples removed from the beginning of each sequence to reduce the recency effect.

The first aim of filtration was to extract and to reject the signals from the subjects, which were giving unreliable scores for most of the time of the experiment. The Spearman rank correlation between each individual average $\bar{y}_{ij..a}$ (2) and mean opinion scores $\bar{y}_{.j..a}$ of p observers (3) was calculated for the whole range of bitrate.

$$\bar{y}_{ij..a} = \frac{1}{n} \sum_{l=1}^{n} \bar{y}_{ijl.a} \qquad (2)$$

$$MOS = \bar{y}_{.j..a} = \frac{1}{p} \sum_{i=1}^{p} \bar{y}_{ij..a} \qquad (3)$$

Signals with weak correlation (less than $0,5$; $\alpha = 0,05$) were discarded (17%). This operation enabled for rejecting data from observers who voted randomly (Fig. 7, observer a) or didn’t respond to changes of quality (Fig. 7 observer b).

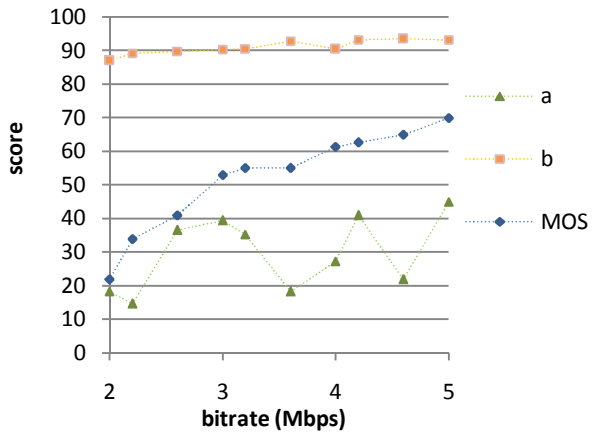

Fig. 7. MOS and two individual averages of scores vs. bitrate.

As mentioned before, mean of scores and the range of scale used vary between observers. Scores from individual observers can be considered as measurement data obtained using instruments with different range and sensitivity (Fig. 8 a). To make particular scores comparable, data normalization was performed (Fig. 8 b).

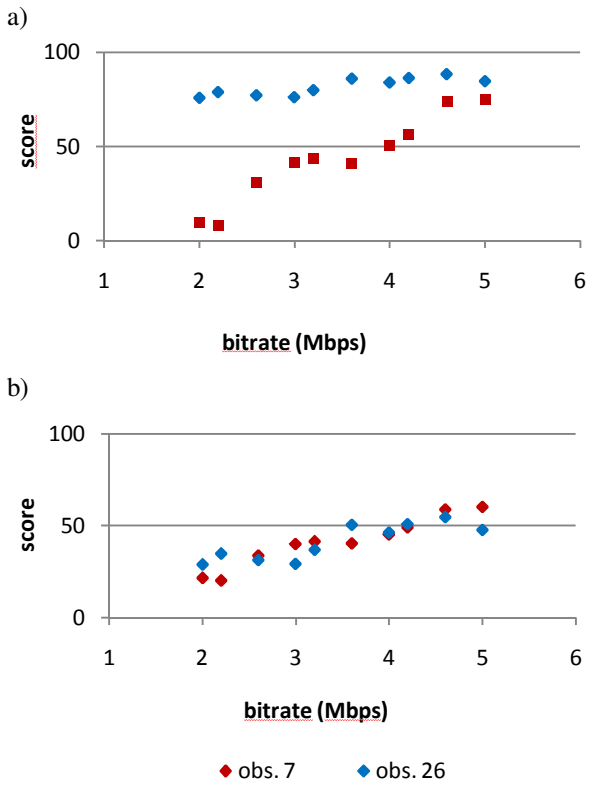

Fig. 8. Two individual average scores vs. bitrate: a) raw data, b) data normalized.

In the next step of data analysis, each level of bitrate was examined separately. The criteria of preserving or rejecting data were based on the differences between:

- scores given by the same observer in replicated assessment of the same sequence – to examine inter-consistency;
- scores given by the observer and the mean opinion scores – to examine intra-consistency.

In order to verify inter-consistency for each *a-th* sequence, *i-th* observer and *j-th* level of bitrate, the Mandel's *k* statistic was computed [17] according to fallowing expression:

$$k_{ija} = \frac{s_{ija}}{s_{rja}} \tag{4}$$

where

s_{ija} - cell standard deviation for *i-th* observer and the *j-th* coding level of *a-th* sequence (5)

s_{rja} - repeatability standard deviation (6).

$$s_{ija} = \sqrt{\frac{\sum_{l=1}^{n}\left(\bar{y}_{ijl.a} - \bar{y}_{ij..a}\right)^2}{n-1}} \tag{5}$$

$$s_{rja} = \sqrt{\frac{\sum_{i=1}^{p} s_{ija}^{\,2}}{p}} \tag{6}$$

The critical values of *k*-statistic can be expressed by the following equation:

$$k_c = \sqrt{\frac{pF\{\alpha, f_1, f_2\}}{F\{\alpha, f_1, f_2\} + (p-1)}} \tag{7}$$

$$f_1 = n-1 \qquad f_2 = (n-1)(p-1)$$

where: α – significance level ($\alpha = 0{,}05$ was assumed)

$F\{\}$ – the inverse of the *F*-distribution with the degrees of freedom f_1 and f_2

If k_{ija} value obtained from experiments exceeded critical value k_c, the scores given by the *i-th* observer for the *j-th* level of *a-th* sequence coding were rejected.

To examine intra-consistency, Mandel's *h* statistic [17] was used. For each *i-th* observer on each *j-th* level of the *a-th* sequence coding, h_{ija} value was computed as follows:

$$h_{ija} = \frac{\bar{y}_{ij..a} - \bar{y}_{.j..a}}{s_{mja}} \tag{8}$$

$$s_{mja} = \sqrt{\frac{\sum_{i=1}^{p}(\bar{y}_{ij..a} - \bar{y}_{.j..a})^2}{p}} \tag{9}$$

where

The critical values of *h*-statistic are expressed by the equation (10):

$$h_c = (p-1)t\{\alpha, f\}\sqrt{p\left(t^2\{\alpha, f\} + p - 2\right)} \tag{10}$$

where $f = p-2$

$t\{\}$ – is the inverse of the two-tailed *t*-distribution with the degree of freedom $f = (p-2)$

If $|h_{ija}|$ exceeded the critical h_c the scores given by the *i-th* observer for the *j-th* coding level of *a-th* sequence were rejected.

Mandel's statistics give more detailed evidence, while they can be computed separately for the individual trueness and precision of one observer compared to the results of all panel of subjects.

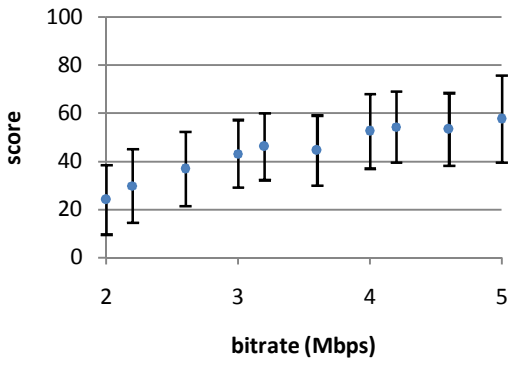

Fig. 9. MOS and standard deviation intervals vs. bitrate, before data processing; sequence 'mobl'; 45 observers.

The final result of proposed data processing was significant decrease in standard deviation of scores given for various bitrate levels and different sequences. from 14 – 18 % (Fig. 9) to 7 - 9 % of the measurement scale (Fig. 10).

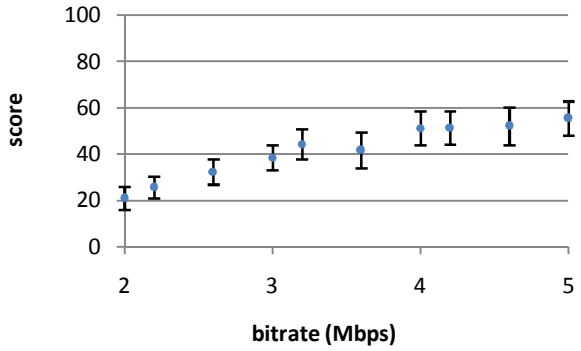

Fig. 10. MOS and standard deviation intervals vs. bitrate, after data processing; sequence 'mobl'.

The MOS signal after the filtration seems to be more sensitive to the quality temporal variations (Fig. 11) but the most important is better consistency of observers' opinions. Smaller standard deviations ensure proportionally narrower confidence intervals. Therefore presented data processing shows promise for all researchers who inquire the information on human visual perception of coding errors both for scientific and practical purposes.

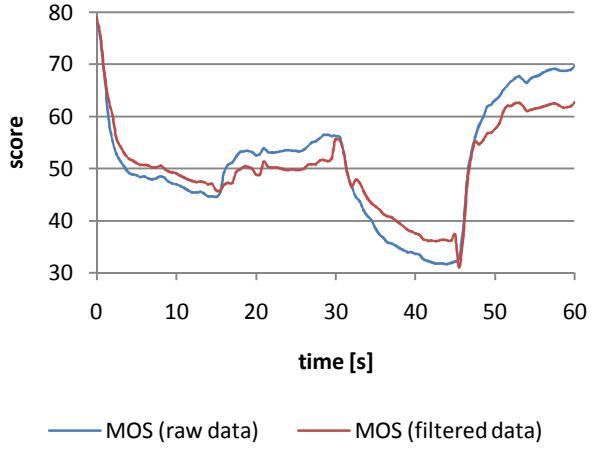

Fig. 11. MOS computed on the basis of all measurement data and on the basis of data after filtration.

4. CONCLUSIONS

The proposed technique for filtration data obtained from observers is a completely new approach. It lets researchers preserve the bigger amount of data by discarding just a part of measurement signals. It enables to decrease standard deviation of all scores by a half. Diminishing of confidence intervals allows examining the influence of numerous factors, such as: age and social background of viewers, observation condition, on the perception and quality demands. And what is the most important – it can facilitate and intensify the development of Human Visual System and consequently create new compression algorithms and video quality analyzers.

ACKNOWLEDGMENTS

This scientific research work was sponsored by the funds for science in years 2007-2009 as research program N N505 4282 33.

REFERENCES

[1] *ITU-T Recommendation* P.910, "Subjective video quality assessment methods for multimedia applications", Apr. 2008.

[2] *ITU-R Recommendation* BT.500-11, "Methodology for the subjective assessment of the Quality of Television Pictures", June 2002.

[3] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies", *SPIE Video Communications and Image Processing Conference*, Lugano, Jul. 8-11 2003.

[4] D. S. Hands, M.D. Brotherton, A. Bourret, D. Bayart: "Subjective quality assessment for objective quality model development", *Electronic Letters*, vol.41, No.7, March 2005.

[5] K. T. Tan, M. Ghanbari: "A combinational automated MPEG video quality assessment model", *Image Processing and its Applications*, Conf. Publication No. 465, IEE 1999.

[6] P. N. Gardiner, M. Ghanbari, D.E. Pearson, K.T. Tan: "Development of a perceptual distortion meter for digital video", *International Broadcasting Convention*, Conference Publication, No. 44712-16, Sept. 1997.

[7] Th. Alpert, J.-P. Evain: "Subjective quality evaluation – The SSCQE and DSCQE methodologies", *EBU Technical Review,* Spring 1997.

[8] D. Abraham, M. Ardito, L. Boch, A. Messina, M. Stroppiana, M. Visca: "Attempts at correlation between DSCQS and objective measurements", *EBU Technical Review,* Spring 1997.

[9] N. Suresh and N. Jayant (USA), "Subjective video quality metrics based on failure statistics" *Circuits, Signals and Systems*, (493), 2005.

[10] Y. Kato and K. Hakozaki, "A video classification method using user perceptive video quality", Proceeding (516) *Internet and Multimedia Systems and Applications*, 2006.

[11] A. Ostaszewska, S. Żebrowska-Łucyk, R. Kłoda, "Metrology properties of human observer in compressed video quality evaluation", *XVIII IMEKO World Congress*, Rio de Janeiro, Sept. 2006.

[12] B. L. Jones, P. R McManus, "Graphic scaling of qualitative terms", *SMPTE Journal*, November 1986.

[13] N. Narita, "Graphic scaling and validity of Japanese descriptive terms used in subjective evaluation tests", *SMPTE Journal*, July 1993.

[14] M. T. Virtanen, N. Gleiss, M. Goldstein, "On the use of evaluative category scales in telecommunications", *Proc. Human Factors in Telecommunications,* 1995.

[15] R. Hamberg, H. Ridder, "*Time varying image quality: modelling the relation between instantaneous and overall quality*", SMPTE Journal, pp. 802-811, Nov. 1999.

[16] Stefan Winkler: "Issues in vision modelling for perceptual video quality assessment", *Signal Processing,* pp. 231-252, 78 (1999).

[17] *ISO 5725-2:1994*, "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".