Annals of Warsaw University of Life Sciences - SGGW
Forestry and Wood Technology № 108, 2019: 97-103
(Ann. WULS - SGGW, For. and Wood Technol. 108, 2019)

The use of the Krippendorff's coefficient in determining intra-rater reliability in human visual quality control of furniture manufacturing processes

KATARZYNA ŚMIETAŃSKA, PIOTR PODZIEWSKI
Department of Mechanical Processing of Wood, Faculty of Wood Technology, Warsaw University of Life Sciences – SGGW

Abstract: *The use of the Krippendorff's Coefficient in determining intra-rater reliability in human visual quality control of furniture manufacturing processes.* The article presents the results of research on the effectiveness of human visual quality control of furniture manufacturing processes. The aim of this experiment was to check the inter-rater reliability of three elements: the judge, the instruction of which the quality of individual objects made of laminated MDF was determined and the quality assessment by a human. The Krippendorff's Coefficient was used as a measure of visual control effectiveness. The results of the experiment showed that the use of human visual quality control of furniture manufacturing processes can be a good solution. The values of all three coefficients K1, K2 and K3 gave alpha> 0.8, which is considered a guarantee of high compliance. Therefore, the tested judge turned out to be reliable and competent, the instructions were clear and sufficient, and the figures obtained were reliable. It turned out that the Krippendorff's Coefficient values differ depending on the number of categories (for two categories K2 = 0.904, K3 = 0.902; for three categories K1 = 0.849). The way the categories were created did not affect the stability of the assessment. The alpha coefficient can undoubtedly be considered a good, convenient measure of monitoring the reliability of the measurement system, because by giving a specific numerical rating, it allows one to control the effectiveness of implemented improvement actions.

Keywords: Krippendorff's Coefficient, intra-rater reliability, human visual quality control, melamine-faced MDF, milling

INTRODUCTION

The aim of quality control of production plants is to check the compliance of the process or product with the requirements of the internal or external customer (Hamrol 2007) by eliminating or minimizing defects and incompatibility in the product. Most often quality control is performed by direct measurement or observation, and its results are data enabling interpretations of the assessed process or product state. In many industries (also in the furniture industry), there is a need for control that will accurately describe the quality of manufactured products, provide information about the process capability and indicate any areas for improvement.

Despite the development of measuring methods based on increasingly objective measuring instruments and the widespread automation of technology (most operations in the manufacturing process are carried out by machines), man is still indispensable, and in many companies the dominant quality control method is organoleptic. Machines collect information in the form of measuring specific product features, but making decisions regarding control activities (dispositions) is a problematic stage. The main reason is the continuous increase in the number and diversity of customer needs, which often requires an unconventional approach to product quality control in the production process or acceptance control. In such cases, the automatic system is not able to ensure an adequate level of repeatability and reproducibility of the inspection and visual inspection carried out by man is the only and most appropriate solution (Hamrol 2007). A man, thanks to their knowledge and experience, is able to respond in a flexible manner, adequate to the situation. In contrast, a machine does not have the ability to learn from experience and improve its work, as well as to improvise or analyse alternative solutions.

Visual inspection is particularly important for processes whose repeatability and reproducibility is limited and the process results differ and require an individual approach when assessing the quality of their performance. The presence of a man in control operations of technological processes is indispensable, especially due to the increasing complexity of manufactured products. Human participation in visual control is valued for its ability to recognize new cases, flexibility in non-standard situations. Making decisions regarding the quality of controlled products requires not only a specific industry knowledge package but often an individual approach to each controlled piece and high sensitivity to incompatibility and limited confidence in the manufactured product.

It should be emphasized that visual quality control carried out by man is characterized by a much higher level of subjectivity, which becomes extremely important in the case of objects and devices of everyday use, such as furniture, interior design elements, cars and household appliances. Only a human is able to subjectively decide which of the elements in the further process of use will satisfy him.

The advantages of visual control (visual inspection) due to which it is often used in production include: simplicity of carrying out in comparison with other types of control, low costs, speed, and a low number of samples. This method does not require specialized measuring equipment, the role of which is played by human vision, and is a non-destructive method. Therefore, manufacturers of products from various industries (automotive (Vogt et al. 2015), electronic (Vogt et al. 2010)) commonly use traditional visual quality control methods, supported only by auxiliary measurements using measuring devices, despite the awareness of the imperfections of the organoleptic method and the significant risk of not detecting an incompatibility or its inappropriate assessment by an employee (controller) (Giesko et al. 2011).

However, you should be aware of the fact that visual control does not guarantee a clear, correct assessment, the primary reason being the limited human reliability. The complexity of the problem of the credibility of a human visual assessment is due to the fact that its effectiveness is influenced by many factors, both organizational and directly related to the human being (Drury et al. 1986; Hamrol et al. 2011). Those factors can be divided into 5 categories (Vogt et al. 2015):

- technical (type of defects, defect visibility, quality level, standards (tests), control automation, other),
- psychophysical (age, sex, observation skills, experience, temperament, creativity, other),
- organizational (training, scope of decision making, feedback, precise instructions, other),
- workplace environment (light, noise, temperature, work time, workstation organization, other), and
- social (team communication, pressure, isolation, other).

There are three sub-types of reliability in literature (Weber 1990; Krippendorff 1980):

1. stability (intra-rater reliability) – consists in re-encoding the same data by the same people; largely refers to the coder's skills,
2. accuracy – measures the compliance of the coding of the material with respect to the standard, established by a group of experts or based on previous research,
3. reproducibility (inter-reliability) – involves checking the degree of consistency in coding the same material by several people; the measurement is based on an estimate of the proportion of consistent categorizations between judges to all their decisions; is defined as the degree of agreement between judges or reliability between judges.

During visual control in industrial practice, the intra-rater reliability becomes the biggest, basic limitation (problem), which is the result of the consistency of several assessments of a given controller. As one knows, products may not be the only the subject of non-compliance in production; it can also be the results of the work of the evaluating

controller. They may result from limited sensitivity to errors, limited perception as well as experience or skills. The causes can be divided into: direct (ignorance or limited opportunities to perform work properly) and indirect (non-ergonomic workstation or incorrect operation of machines). It is, therefore, necessary to regularly check the effectiveness of human visual quality control in manufacturing plants. The appearance of conformity assessment errors in the form of incorrect product classification may contribute to deterioration of process efficiency.

The basic, most frequently used tools for measuring reliability are (Lombard et al. 2004, Krejtz et al. 2005, Krippendorf 2004, Neuendorf 2002, Scott 1955, Rosenfield 1986, Fleiss 1971, Cohen 1960, Hayes 2007, Holsti 1969):

- joint-probability of agreement (percent agreement),
- Holsti's method (Holsti's CR),
-Scott's (1955) π (pi),
- Cohen's κ (Kappa),
- Krippendorff's α,
- Kendall's W.

Therefore, quality systems in industry should set themselves the task of not only focusing on finding and eliminating the root causes of errors but also monitoring the effectiveness of the work of raters. This will allow creating a reliable source of feedback on the production process, which will improve production and reduce manufacturing costs.

These tools allow you to assess the capabilities of a measuring system consisting of quality controllers. They are a good solution for monitoring the reliability of the measurement system because by giving a specific numerical rating, it allows you to assess the effectiveness of implemented improvement actions.

MATERIALS AND METHODS

The research used Krippendorff's Coefficient – the coefficient of agreement between individual assessments as the most appropriate statistic for assessing intra-rater reliability in human visual quality control. It allows to determine the degree of credibility (compliance) of the results obtained between multiple repetitions of the same set of objects assessment by one observer, giving the repeatability and credibility of the visual assessment of one judge.

Krippendorff's alpha allows uniform reliability standards to be applied to a great diversity of data (Krippendorff 2004, Krippendorff 2011):

- it can be used for any number of values per variable
- can be used for any number of observers
- can be used for small and large samples small and large sample sizes
- to several metrics (scales of measurements) – nominal, ordinal, interval, ratio, and more
- applies to data sets with missing values to data with missing values

The coefficient is defined in the simplest form [1]:

$$\alpha = 1\text{-}D_o/D_e \qquad [1]$$

Whrere:

D_o – a measure of the observed disagreement

D_e – a measure of the disagreement that can be expected when chance prevails

When agreement is observed to be perfect and disagreement is, therefore, absent, Do = 0 and α = 1, indicating perfect reliability.

When agreement and disagreement are a matter of chance and observed and expected disagreements are equal, De = Do and α = 0, indicating the absence of reliability.

The assessment procedure required special preparation of test samples and appropriate conditions. Samples were elements made of a MDF laminated board (one of the basic materials in the furniture industry). In each element, a groove and a step were milled, forming three edges for the evaluation of machining quality (k1, k2, k3). The dimensions and shape of the object are shown in Fig. 1. All assessments always take place in the same room, during one uninterrupted session, with similar lighting and sample presentation (all elements were displayed simultaneously). During the tests, the judge (under time pressure, which allowed to simulate working conditions in industrial conditions), evaluated the subject edges four times (each separately) on a scale of 1–3. The following is the content of the instructions:

1. The judge assesses not the quality of the whole item but the quality of each of the 3 indicated edges separately (each separately). The edges are marked as k1, k2, k3 (as shown in the drawing below).

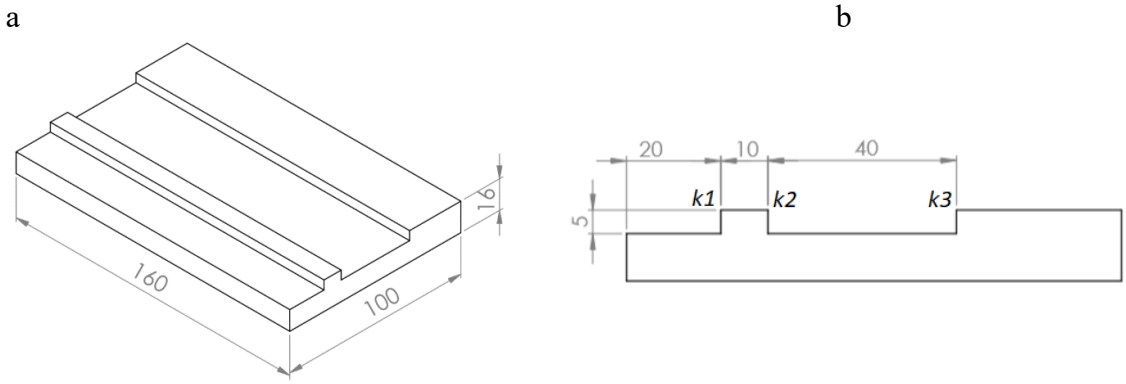

Figure 1. The shape of the element (a) and the location of the individual edges to be assessed (b).

2. The judge should not pay attention to the condition of the corners (short zones at the beginning and end of each edge). Possible damages on the first and last 5 mm edges may have non-processing reasons and as such should not affect the quality assessment of the entire edge. The length of ignored zones (5 mm) is given with a significant margin and can be treated as an estimate (you do not need to use a ruler, so-called "eye-balling" is sufficient).

3. The judge should not attempt to pretend to be a professional (e.g. factory) quality controller but be guided by his own "intuition" (as if he were an ordinary customer who watches the goods in a store).

4. The judge has three ratings to choose from:

– rating 1 = very good, good or at least satisfactory quality without any major "aesthetic discomfort"

– rating 2 = quality creating some moderate "aesthetic discomfort" but conditionally acceptable, e.g. if it was less visible edges of the furniture or in the case of a very attractive product price

– rating 3 = clearly unambiguous quality, definitely unacceptable from an "aesthetic point of view".

RESULTS AND DISCUSSION

The obtained numerical data (assessments of the "JD" judge) were recorded in a tabular form. Based on the results of the conducted tests, three values of the measurable Krippendorff's numerical coefficient α were determined as follows (without distinguishing between individual edges k1, k2, k3):

- K1 – the value of Krippendorff's Coefficient was calculated for a three-grade evaluation – the quality of the elements was assessed on the basis of three categories (three grades 1, 2, 3, in accordance with the instructions)
- K2 – Krippendorff's Coefficient value was calculated for a two-stage evaluation – the results of two better categories (grades 1 and 2 in accordance with the instructions) were combined into one
- K3 – Krippendorff's Coefficient value was calculated for a two-stage evaluation – the results of two worse categories (grades 2 and 3 according to the instructions) were combined into one

The obtained results are presented on the graph (Fig. 2).

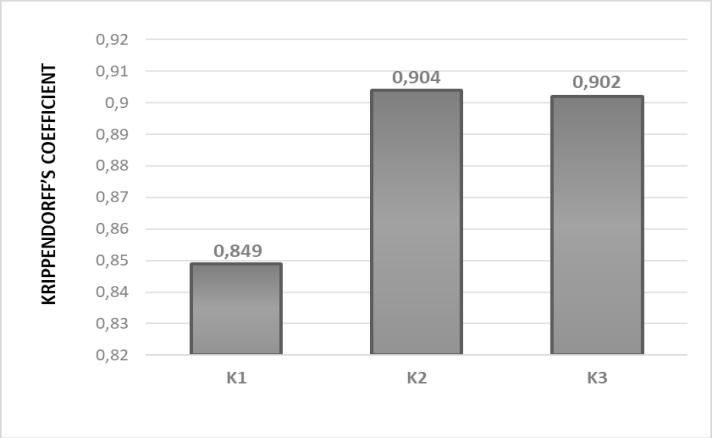

Figure 2. Krippendorff's Coefficient values for different assessment variants

CONCLUSIONS

1. The use of human visual quality control of furniture manufacturing processes proved to be a good solution. The values of all three alpha coefficients (for K1, K2 and K3) exceeded 0.8, which is considered a guarantee of high compliance. Therefore, the JD judge turned out to be reliable and competent, and the figures obtained should be considered reliable.

2. The number of categories played a significant role. For two categories, Krippendorff's Coefficient takes higher values (K2 = 0.904, K3 = 0.902) than for three categories (K1 = 0.849). The way categories are created does not significantly affect the stability of the rating.

3. The alpha coefficient can undoubtedly be considered a good, convenient measure of monitoring the reliability of the measurement system because, by giving a specific numerical rating, it allows to control the effectiveness of implemented improvement actions.

REFERENCES

1. COHEN, J. (1960). "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement* 20, pp. 37–46.
2. DRURY, C.G., KARWAN, M., VANDERVARKER, D.R. (1986). "The Two-Inspector Problem", *IIE Transactions* 18(2), pp. 174–181.
3. FLEISS, J. L. (1971). "Measuring nominal scale agreement among many raters", *Psychological Bulletin* 76(5), pp. 378–382.
4. GIESKO, T., MAZURKIEWICZ, A., ZBROWSKI, A. (2011). "Optomechatroniczny system do automatycznej kontroli jakości wyrobów w przemyśle", *Problemy eksploatacji* 4, pp. 103–114 (http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.baztech-article-BAR0-00600085 pdf])
5. HAMROL, A. (2007). Zarządzanie jakością z przykładami. PWN, Warszawa
6. HAMROL, A., KOWALIK, D., KUJAWIŃSKA, A. (2011). "Impact of Selected Work Condition Factors on Quality of Manual Assembly", *Human Factors and Ergonomics In Manufacturing* 21(2), pp. 156–162.
7. HAYES, A.F., KRIPPENDORFF, K. (2007). "Answering the Call for a Standard Reliability Measure for Coding Data", *Communication Methods and Measures* 1(1), pp. 77–89.
8. HOLSTI, O. R. (1969). Content analysis for the social sciences and humanities. Reading, MA: AddisonWesley
9. KREJTZ, K., KREJTZ, I. (2005). "Rzetelność w analizie treści", pp. 217–230 in: Stemplewska-Żakowicz, K., Krejtz, K. (red.) Wywiad psychologiczny. Wywiad jako postępowanie badawcze (2005), Pracownia testów psychologicznych polskiego towarzystwa psychologicznego, Warszawa
10. KREJTZ, K., KREJTZ, I. (2005). "Wybrane statystyki zgodności miedzy sędziami w analizie treści", pp. 231–249 In: Stemplewska-Żakowicz K., Krejtz K. (red.) Wywiad psychologiczny. Wywiad jako postępowanie badawcze (2005), Pracownia testów psychologicznych polskiego towarzystwa psychologicznego, Warszawa
11. KRIPPENDORff, K. (2004). "Reliability in Content Analysis: Some Common Misconceptions and Recommendations", *Human Communication Research* 30(3), pp. 411–433.
12. KRIPPENDORFF, K. (2004). Content analysis: An introduction to its methodology. Second Edition. Thousand Oaks, London, New Delhi: SAGE Publications
13. KRIPPENDORFF, K. (2011). "Computing Krippendorff's Alpha-Reliability", fromhttp://repository.upenn.edu/asc_papers/43
14. KRIPPENDORFF, K. (1980). "Validity in Content Analysis", *Computerstrategien fur die Kommunikationsanalyse* 3, pp. 69–112
15. LOMBARD, M., SNYDER-DUCH, J., & BRACKEN, C.C. (2002). "Content analysis in mass communication research: An assessment and reporting of intercoder reliability", *Human Communication Research* 28, pp. 587–604.
16. NEUENDORF, K. A. (2002). The content analysis guidebook. SAGE Publications, Thousand Oaks
17. SCOTT, W.A. (1955). "Reliability of content analysis: The case of nominal scale coding", *Public Opinion Quarterly* 19(3), pp. 321–325.
18. SCOTT, W. A. (1955). "Scott's π (Pi). Reliability for nominal scale coding", pp. 347–349 In: Krippendorff, K., Bock, M.A. The Content Analysis Reader (2009), Los Angeles, London, New Delhi, Singapore

19. ROSENFIELD, G.H., FITZPATRICK-LINS, K. (1986). “A coefficient of agreement as a measure of thematic classification accuracy”,*Photogrammetric engineering and remote sensing* 52(2), pp. 223–227.
20. WEBER, R. P. (1990). Basic content analysis. SAGE Publications, Newbury Park, Londorn, New Delhi
21. VOGT, K., KUJAWIŃSKA, K. (2015). “Human factors in visual quality control”, *Management and Production Engineering Review* 6(2), pp. 25–31.
22. VOGT, K., KUJAWIŃSKA, K. (2010). “Analiza wpływu wybranych czynników pracy na skuteczność kontroli wzrokowej”, *Inżynieria Maszyn* 18(1), pp. 40–50.

Streszczenie: *Wykorzystanie współczynnika Krippendorffa do określania wiarygodności oceny ludzkiej w wizualnej kontroli jakości procesów produkcji mebli*. W artykule zaprezentowano wyniki badań skuteczności ludzkiej wizualnej kontroli jakości procesów produkcji mebli. Eksperyment miał na celu sprawdzenie wiarygodności ocen wg. trzech elementów: wiarygodności oceniającego sędziego, instrukcji na podstawie której określano jakość poszczególnych przedmiotów wykonanych z laminowanej płyty MDF oraz oceny jakości przez człowieka. Jako miarę skuteczności kontroli wizualnej zastosowano współczynnik Krippendorffa. Wyniki eksperymentu pozwolił stwierdzić, że zastosowanie ludzkiej wizualnej kontroli jakości procesów produkcji mebli może być dobrym rozwiązaniem. Wartości wszystkich trzech współczynników K1, K2 oraz K3 dały alpha > 0,8, co uznaje się za gwarancję wysokiej zgodności. Sędzia JD okazał się, więc osobą rzetelną i kompetentną, zaproponowana instrukcja jasna i wystarczająca, a otrzymane dane liczbowe wiarygodne. Okazało się, że wartości Krippendorff’s Coefficient różnią się w zależności od ilości kategorii (dla dwóch kategorii K2=0,904, K3=0,902; dla trzech kategorii K1=0,849). Sposób tworzenia kategorii nie wpływał na stabilność oceny. Współczynnik alpha można niewątpliwie uznać za dobrą, wygodną miarę monitorowania wiarygodności systemu pomiarowego, gdyż dając konkretną ocenę liczbową, pozwala kontrolować skuteczność wprowadzanych działań doskonalących

Corresponding authors:

Katarzyna Śmietańska, Piotr Podziewski,
Faculty of Wood Technology SGGW,
Department of Mechanical Processing of Wood,
ul. Nowoursynowska 159,
02-776 Warsaw,
Poland
e-mail: katarzyna_smietanska@sggw.pl
e-mail: piotr_podziewski@sggw.pl

ORCID ID:
Śmietańska Katarzyna 0000-0001-8705-3700
Podziewski Piotr 0000-0002-2628-5062