ANALYSIS OF ASSESSMENT OF STUDENTS IN SUBJECTS THAT ARE COMPLETED WITH AN EXAM IN MASTER OF PUBLIC HEALTH PROGRAM

M. Panczyk, A. Zarzeka, J. Belowska, Ł. Samoliński, J. Gotlib

Division of Teaching and Outcomes of Education, Faculty of Health Sciences, Medical University of Warsaw (POLAND)

Abstract

Introduction

Educational diagnostics on university level is currently based on evaluation of a student's progress with particular attention placed on measuring the achieved learning outcomes.

In case of Public Health Faculty, we are faced with a broad and interdisciplinary field of skills and knowledge. A process of education that has been carried out properly and comprehensive assessment of the gained competences is the basis for further research into the learning outcomes and it also allows to analyse the fate of graduates of a given major. Evidence-based assessment should be applied as good practice at any university which attempts to adjust their educational policy to the changing needs and requirements of labour market.

Aim of study

Analysis of assessment in the subjects that are completed with an exam on the second degree studies at the faculty of Public Health.

Materials and Methods

Examination data of 565 students from the faculty of Public Health who qualified to the studies (age average 22.4 ± 1.22) concerned those who began their second degree of studies at the Faculty of Health Sciences at the Medical University of Warsaw (MUW) between the years 2007-2012. Learning outcomes in nine subjects finished with an exam were analysed retrospectively and these were the following: *Biostatistics, Public Health in Practice, Advances in Health Promotion, Economics, Forms of Health Care, Organisation and Management in Health Care, Financing in Health Care, Health Care Law*, *Epidemiology*.

In order to evaluate intercorrelation of learning outcomes for individual subjects, r-Pearson's linear correlation coefficient was used. In order to determine trends in assessing students in consecutive years, Kruskal-Wallis' non-parametre ANOVA test of ranges was applied as well as Leven's homogeneity of variance test. For comparative analysis of dependent variables, Friedman's non-parametre test was used with Kendall's coefficient of concordance.

For all analyses, the a priori level of significance was established at 0.05.

Results

Evaluation of intercorrelation shows that for eight analysed subjects, there are positive intercorrelations between students' learning outcomes in individual areas of Public Health. r-Pearson's correlation coefficient remained within 0.09 and 0.44. Only in the case of *Public Health in Practice* there was no statistical differences in intercorrelation in the results that students obtained in two subjects (*Organisation and Management in Health Care* and *Health Care Law*), and where relevance was maintained, r-Pearson's coefficients were significantly lower than for other analysed subjects.

Analysis of internal consistency in assessing students in individual subjects in consecutive years showed that for none of the studied subjects homogeneity was observed when analysing students' achievements (no homogeneity of variance, Leven's test, $P< 0.001$ and ANOVA test of Kruskal-Wallis' ranges, $P< 0.001$). Moreover, combined comparative analysis of average marks obtained by students in individual subjects between the years 2007-2012 shows that the area of knowledge that was worst-rated, was *Biostatistics* (average of marks 3.0 ± 0.55), whereas the best-rated one was *Forms of Health Care* (average of marks 4.3 ± 0.49). For the rest of subjects, students' average marks remained within the range of 3.6 and 4.0, which points out to an average level of concordance (Kendall's concordance coefficient 0.25).

ISBN: 978-84-606-5763-7

Conclusions

The observed lack of homogeneity in evaluating students' achievements is the evidence of insufficient consistency in measuring learning outcomes. Critical judgement of currently used solutions allows to plan detailed evaluation of applied assessment criteria so as to diagnose the weak points in the teaching areas and evaluate reliability and accuracy of the used methods that assess competences of students at the Faculty of Public Health.

Keywords: performance evaluation system, performance assessment, achievement assessment, educational diagnosis, educational measurement.

1 INTRODUCTION

Public Health is an interdisciplinary direction of studies that is on the border on health, social and medicine sciences as well as physical education.

This subject is taught to students of both 1st and 2nd degree. A graduate of the first should have knowledge, skills and social competences concerning health prevention, population health, human organism functioning, epidemiology and sanitary and epidemiological supervision as well as health promotion, regulations concerning nourishment and the basis of economy.

Studies of the 2nd degree are of such character that vastly broadens knowledge and skills in the subject matter. Education includes advanced courses in economy, management, epidemiology, biostatics, methodology of scientific research or health promotion, additionally it introduces relevant elements of practice in the field of public health. A graduate of the 2nd degree is also equipped with knowledge concerning health care law, Evidence-based Health Policy and management in health care. As part of their studies, they also are taught proper and efficient identification of social and environmental health conditioning as well as mutual dependencies that occur between human health, condition of natural environment, system of social care and the socio-economic situation of the state.

Classes in subjects realised as part of the 2nd degree studies are performed in form of lectures, seminars and exercises. For some subjects, methods such as lectures predominate. This concerns for instance health sociology or social politics. Other courses abound in exercises that are performed in small groups (around 10 people). As examples, biostatics or methodology of scientific research can be pointed out. Within these subjects, students can also realise their own research projects and they subject them to statistical analysis.

Due to interdisciplinary character of education and the broad area of knowledge and skills that are required of students of Public Health, appropriate measurement of their achievements is crucial, so that the result of such a measurement be adequate to the level of a graduate's preparation to undertake professional work. Such a measurement of educational achievements, called educational measurement, means assessment of knowledge and skills a student represents that is based on clearly stated rules that can also be experimentally confirmed. However, the rule of non-intuitive and experimentally confirmed educational measurement is often difficult when it comes to realisation. One of the important reasons for this is provided by Van der Vleuten et al. in their publication "*The need for evidence in education"*: "using intuition in [medical] education (...) is, to a large extent, based on ignoring [teachers]"*.* As the authors state, a good academic teacher means something more than just an expert in a given field of science. A professional is a person who bases his educational activities on understanding the issue of *learning and teaching* process [1]. We must be critical in evaluating which practices based on tradition / intuition are valuable and thus eliminating these which are wrong and should not be used in education. Achieving this should not be possible without becoming familiar with theoretical frames that are the basis in assessing achievements that students make. The possibility to compare an educational situation that an academic teacher faces in his professional life with the results and conclusions published in journals dealing with the issue of research into education, is one of the key elements of *Evidence-based medical education/teaching* [2].

2 AIM OF STUDY

The aim of the results of studies presented here is the analysis of coherence in assessing students achievements in subjects that finish in an exam and are included in the curriculum at the second degree of studies at Faculty of Health Sciences, Medical University of Warsaw (MUW).

3 MATERIALS AND METHODS

565 students qualified to the study from the Faculty of Public Health (age mean 22.4 ± 1.22), who undertook full time studies of the second degree at MUW between the years 2007-2012. Characteristics of the studied group is presented in Table 1.

Table 1. Characteristics of the studied group of students at the Faculty of Public Health, who undertook studies at MUW between the years 2007-2012

Class	N	Women	Men	MUW graduates	Other than MUW graduates
2007/08	100	80	20	90	10
2008/09	103	87	16	86	17
2009/10	103	88	15	88	15
2010/11	87	74	13	75	12
2011/12	80	73	7	66	14
2012/13	92	82	10	73	19

Retrospective analysis was used when analyzing the results of teaching in nine subjects that finished in an exam: Biostatistics, Public Health in Practice, Advances in Health Promotion, Economics, Forms of Health Care, Organisation and Management in Health Care, Financing in Health Care, Health Care Law, and Epidemiology.

In order to determine the trends in assessing students in consecutive years, a Kruskal-Wallis one-way analysis of variance and homogeneity of variance Leven's test. So as to perform compatibility assessment for individual students for subsequent exam subjects, a non-parametric Friedman's ANOVA test was used in order to compare dependent variables and Kendall's coefficient of concordance. The analysis of assessment's reliability was determined using α-Cronbach coefficient (Kuder-Richardson coefficient for a test consisting of two-category items). Inter-correlation evaluation for the nine analysed subjects was performed using r-Pearson linear correlation coefficient.

For all analyses, the a priori level of significance was established at 0.05.

4 RESULTS

The analysis of internal consistency in assessing students in individual subjects in consecutive years showed that for none of the studied subjects homogeneity was observed in assessing students' achievements. A high degree of variation in students' grades for individual subjects is confirmed by lack of homogeneity in variance (Leven's test, $P < 0.01$). Moreover, it was proved that there is a statistically relevant difference in assessing students in individual subjects for subsequent years of students (Kruskal-Wallis one-way analysis of variance, $P < 0.01$). Summary of the results of analysis of the degree of diversification of student assessment is presented in Table 2.

Table 2. Summary of the results of analysis of the degree of diversification of student assessment in individual subjects that finish in an exam at the Faculty of Public Health between the years 2007-2012

Subject	Value of the test statistic F	*P*-value *	Value of the test statistic H	*P*-value **
1. Biostatistics	16.44337	< 0.00001	77.75243	< 0.00001
2. Public Health in Practice	8.58829	< 0.00001	60.61327	< 0.00001
3. Advances in Health Promotion	3.07966	0.01	21.34413	0.0007
4. Economics	7.09571	0.000002	20.39947	0.001
5. Forms of Health Care	23.63363	< 0.00001	42.94392	< 0.00001
6. Organisation and Management in Health Care	8.10552	< 0.00001	112.9505	< 0.0001
7. Financing in Health Care	10.63456	< 0.00001	80.37607	< 0.00001
8. Health Care Law	3.25442	0.01	152.3687	< 0.0001
9. Epidemiology	8.32245	< 0.00001	15.18932	0.01

* homogeneity of variance Leven's test

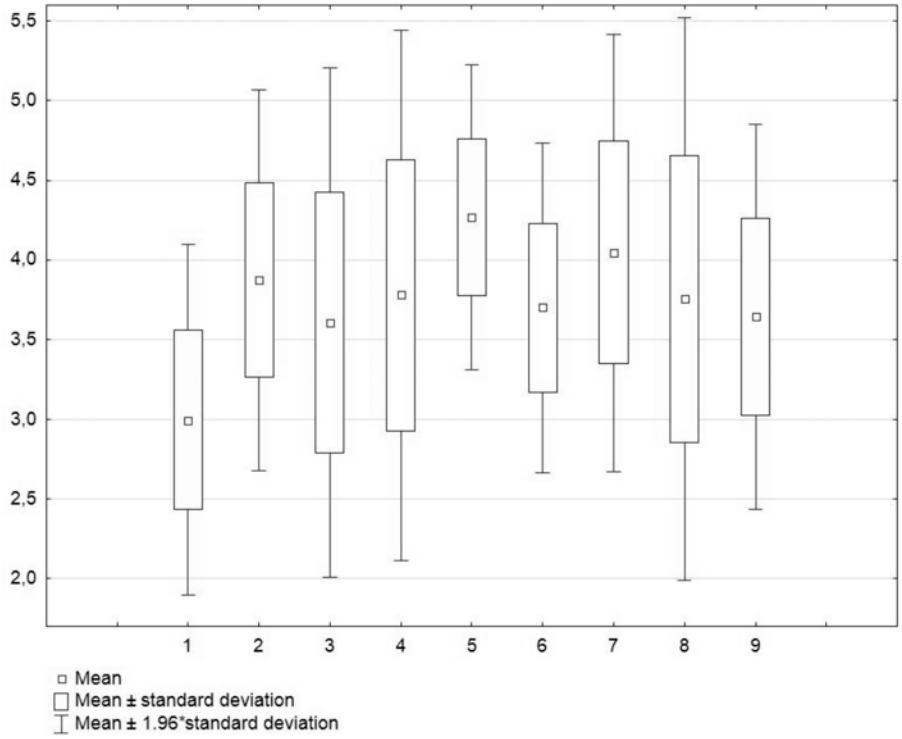
** Kruskal-Wallis one-way analysis of variance

Overall comparative analysis of average grades achieved by the students between 2007-2012 in individual subjects shows that the subject with worst assessment *Biostatistics* (average 3.0 ± 0.55), whereas the best one was *Forms of Health Care* (average 4.3 ± 0.49). For other subjects, students' average grades remained within the range of 3.6 and 4.0. The degree of differentiation in students' achievements measured by the average of their grades for subjects included in the curriculum at the faculty of Public Health between 2007-12 is presented in Fig. 1.

While comparing the grades obtained by a given student in subjects that included education in the field of Public Health, it was found that concordance of educational measurement oscillates around the level that is below the average (Kendall's concordance coefficient in the range between 0.22 and 0.43). Moreover, the analysis of students' achievements in various subjects that was performed using a non-parametric ANOVA Friedman's test (comparison for nine connected groups) indicates some statistically relevant deviation in consistency of assessment ($P < 0.000001$). A detailed summary of the concordance analysis is presented in Table 3.

The method that is most frequently used when evaluating the internal concordance of the results of measurements for at least two factors is an α-Cronbach reliability coefficient. The analysis showed an average level of assessment reliability for all subjects – coefficient $\alpha = 0.74$ with the level of reliability in consecutive years oscillating around 0.67 and 0.81. The detailed evaluation allowed to establish the fact that subject of *Public Health in Practice* (between the years of 2007-2009) and *Forms of Health Care* (between the years of 2010-2012) negatively influence the general concordance of the measurement.

Evaluating accuracy in the inter-correlation analysis shows that for eight of the studied subjects, there are positive dependencies between the results of teaching in individual areas of Public Health. r-Pearson correlation coefficient remained between 0.009 and 0.44. Only in case of *Public Health in Practice* is was found that there were no statistically relevant inter-correlations with the results of teaching obtained in two subjects (*Organisation and Management in Health Care* and *Health Care Law*), and where the relevance remained, r-Pearson coefficients were noticeably lowered than for other subjects that were analysed. On the other hand, the strongest correlations were observed for pairs of subjects *Organisation and Management in Health Care – Health Care Law* ($r = 0.44$) and *Financing in Health Care – Health Care Law* ($r = 0.42$). Summary of the correlation analysis is presented in Table 4.



1 - Biostatistics, 2 - Public Health in Practice, 3 - Advances in Health Promotion, 4 - Economics, 5 - Forms of Health Care, 6 - Financing in Health Care, 7 - Organisation and Management in Health Care, 8 - Health Care Law, 9 – Epidemiology

Figure 1. Structure of students' evaluation in individual subjects that finished in an exam at the faculty of Public Health between the years of 2007-2012

Table 3. Results of comparative analysis of grades in individual students for subjects that finished in an exam at the faculty of Public Health between the years of 2007-2012

Subject	Rank average	Rank sum	Grade point average	Standard deviation	Value of the test statistic (*P*-value *)
1. Biostatistics	2.21	1232.0	2.99	0.5536	1131.481 (< 0.000001)
2. Public Health in Practice	5.44	3035.0	3.87	0.6123	
3. Advances in Health Promotion	4.63	2582.5	3.61	0.8166	
4. Economics	5.08	2832.5	3.78	0.8494	
5. Forms of Health Care	7.00	3903.5	4.27	0.4909	
6. Financing in Health Care	4.80	2680.0	3.70	0.5275	
7. Organisation and Management in Health Protection	6.13	3421.5	4.04	0.7004	
8. Health Care Law	5.15	2872.0	3.76	0.9046	
9. Epidemiology	4.57	2551.0	3.65	0.6176	

* non-parametric ANOVA Friedman's test for comparing the dependent variables

Table 4. Results of inter-correlation analysis for grades of students in individual subjects that finished in an exam at the faculty of Public Health between the years of 2007-2012

	1	2	3	4	5	6	7	8	9
1. Biostatistics	1.00	0.14	0.26	0.34	0.09	0.32	0.34	0.39	0.22
2. Public Health in Practice		1.00	0.21	0.18	0.09	0.03*	0.09	0.07*	0.11
3. Advances in Health Promotion			1.00	0.31	0.18	0.25	0.26	0.29	0.30
4. Economics				1.00	0.35	0.34	0.30	0.39	0.25
5. Forms of Health Care					1.00	0.16	0.13	0.21	0.22
6. Organisation and Management in Health Protection						1.00	0.33	0.44	0.26
7. Financing in Health Care							1.00	0.42	0.25
8. Health Care Law								1.00	0.25
9. Epidemiology									1.00

* $P > 0.05$ (statistically insignificant)

5 DISCUSSION

The issue of evaluating medical education and educating specialists in the field of health sciences is a subject of numerous studies and arises great interest in academic circles. This comes as no surprise as evaluating students is perceived as one of the most important elements in entire education system. On the one hand it describes the level of education that was assumed for a student, on the other it may become a measurement of the quality of education process [3]. However, despite the strong theoretical background, numerous observations and results of studies into evaluation, the concept of academic practice based on dogmas, convictions or intuition and traditions of a given academy is deeply rooted [4]. Such conduct is frequently the reason of failures and also may lead to waste of power and resources for actions that will not bring satisfactory results in form of increasing the quality of education. Such conjuncture is well reflected in the words of George Santayan: "*Those who do not learn from history are doomed to repeat it*" [5]. One needs to bear in mind that the results of studies into education often bring evidence that is contradictory to the teachers' intuition or the existing dogmas.

Regardless of the aim that evaluation serves, it always is connected with a more or less systematic collection of observation data that lead to drawing conclusions concerning their features and properties of a students who undergoes evaluation [3]. For this process to be a highly objective source of information concerning the results of education, it must fulfil certain criteria described as features of educational diagnosis. These indispensable properties of educational measurement may be presented in hierarchical form. Characteristics of lower order and quality criteria that correspond to them are subjects to characteristics of higher order. Among quality criteria, the following can be listed in order from the lowest to the highest: independence of a measured situation, objectivity of scoring, reliability, validity and objectivity of measurement. While performing educational measurement, one needs to remember about growing conventionality, specificity and complexity of individual steps in hierarchy ladder. The quality of measurement always depends on the applied methods of evaluation of the achieved teaching results, but also on external factors that may relevantly influence its properties. Thus obtaining an ideal measurement that would possess all necessary properties to the highest extent is practically impossible [6].

Measurement impartiality, i.e. independence of measurement means creating equal and fair conditions for all students in order to evaluate their achievements. Whereas assessment bias that is frequently the source of systematic errors in measurement leads to unfairly high or low results obtained by a given group of students. Equal treatment of all students in consecutive years means independent assessment of their achievements regardless of the results these students achieved in previous educational cycles, schools / academies they previously attended or students group they

learnt in. A relevant element of measurement impartiality is creating appropriate conditions during examining and adopting such methods of evaluating students' achievements that would provide a comparable level of independence of a measured situation in consecutive years [7]. As demonstrated by the results of consistency of students' assessment in individual subjects in consecutive years, we may claim that there is a high differentiation in evaluation (lack of homogeneity of variance). Moreover, statistically relevant differences in assessing students in a given subject were observed for students of consecutive years. The above results point to the fact that there are factors which influence the degree of differentiation in assessment of students. It is highly unlikely that differences relevant in the structure of grades measured by an average / median and a variance are caused by a step change of the achieved results. Therefore the question whether no repetition in assessment of students is connected with assessment bias or improper evaluation of educational measurement methods is crucial, or perhaps it is their poor quality, remains open.

What is directly connected with measurement impartiality, is accuracy of scoring (objective scoring) understood as adequacy of the measurement scale to the evaluated properties. Practically, achieving a high accuracy in scoring on the one hand depends on the manner in which test questions were created and the quality of rubrics concerning assessment described in the key (construction causes), and on the other is a derivative of competences, professional experience and examiner's personality (personal causes) [8]. The most frequent source of discrepancies in relation to the results of measurement is too great severity / leniency of the examiner and the tendency to give extreme grades or excessive averaging [9]. As can be seen from the results of analysis concerning grades in individual subjects, in case of *Biostatistics* students obtained noticeably lower grades in comparison with other subjects (a severe examiner), whereas for *Forms of Health Care* these grades were extremely high (a too lenient examiner). Moreover, for the three analysed subjects (*Advances in Health Promotion, Economics, Health Care Law*), a wide range of results inter-changeability was observed, which naturally reflects the differentiation in achievements in the studied group of students. Extreme values do not have to, of course, mean lack of objectivity in scoring, but they may result from the fact that a given subject needs complex competences that are difficult to master (a very low score average in students) or even on the contrary – achieving results in education within a given range is relatively easy for students (a high average of grades). A differentiated level of expectations towards students in individual subjects may, to certain extent, justify the result of analysis of their achievements, pointing to statistically relevant deviation in the consistency of assessment. However, consistency of assessment differed in consecutive years. It should be expected that regardless of the year, a good student, as opposed to the weak one, will have higher grades in all subjects. A measurement of such assessment consistency is Kendall Tau rank correlation coefficient which is based on the difference between the probability of the fact that two variables are placed in the same order within the observed data, and the probability of the fact that their order differs [10]. Value of this coefficient was very different for consecutive years (between 0.22 and 0.43), which is evidence that in all cases the rules of objectivity of score while assessing students in consecutive years were not maintained. Errors in the range of scoring accuracy may be connected with excessive rigidity in the criteria of assessment in a situation where scoring is introduced not for educational reasons but results solely from the appropriateness of solving a given task, which does not necessarily fit into the range of knowledge and skills assumed for the educational effect being measured. This issue will be of particular importance when assessing reliability and validity of measurement [6].

Measurement reliability is recurrence of obtained results in certain conditions. If these results are the same or very similar in established circumstances (e.g. in exam situation), then such educational measurement may be considered a reliable one. In relation to social studies, Earl Babbie pointed to a necessary condition that determines reliability: impartiality of measurement conditions and precision in scoring [11]. Insufficient reliability of the applied procedure while assessing drastically contributes to a lower level of trust, considering the fact that in similar circumstances individual results differ significantly from each other. To assess whether a given measurement is reliable, various analytical methods may be applied. The most frequently used one is determining the degree of correlation of individual exams or their fragments (e.g. odd-even or split-half reliability) [12], or internal consistency assessment of the results of measurement by assessing average variances for all exams (α-Cronbach – 20 Kuder-Richardson coefficient for a test comprising of two-category items) [13]. As is presented in the included results of reliability measurement performed using α-Cronbach, assessment of students' achievements carried out between the years 2007 – 2012 was characterised by a sufficient level of reliability ($\alpha = 0.74$). However, the analysis of individual years shows that in consecutive years the level of internal consistency varied (α from 0.67 to 0.81). Inappropriate selection of exam methods and inappropriate construction of exam tasks in particular, which are the basis of assessment, cause

lowering the measurement's reliability. A student may not have the opportunity to fully present his / her achievements in a given field if an exam radically narrow the content that aims at assessing educational outcomes. This problem may concern two subjects in particular, i.e. *Public Health in Practice a*nd *Forms of Health Care*, for which negative influence was observed when assessing the overall consistency of measurement. Another reason for the low value of α coefficient (< 0.7) could be a higher percentage of random errors in the results of a given measurement. Random fluctuations in results of assessment described in a classic theory of an exam test may lower the value of α coefficient [14]. With α values = 0.5, random errors constitute as much as a half of variation of the obtained results, and a measurement performed in such conditions may be applied solely during inter group comparisons and not while performing individual differentiation. It also needs to be pointed out that mere aiming at obtaining high values of α coefficient does not rectify the problem of reliability since a high α value only means minimizing the influence of random errors on the obtained results, yet it does not provide assurance that systematic errors, sometimes serious ones, will not occur [15].

Apart from studies that focus on the issue of reliability that refer to the question "how measuring should be performed?", determining validity of measurement is also an important factor while creating good assessing tools, which would answer the question "what is being measured?". Validity in this area needs to be understood as a degree of consistency that a measuring tool measures in order to measure the item it was designed to for. Thus, it is the usefulness of a given method in assessing a specific set of features and properties of an examinee [16]. If the selected method tests the skills of a student's ability to adjust to an applied measuring tool (literally "What Do They Want Me To Say?"), then this assessment is not directed at these features that we wish to measure [17]. There is no exact method of measuring validity, but just its average evaluation, which is usually based on applying one of the three concepts according to which validity of measurement can be determined and these are: content [18], empirical [19], and construct validity. [20]. One important type of validity is content validity – the degree of correspondence between the contents of the exam and the logical and curricular domains intended to be measured [21]. In case of exams that are included in the curriculum at the faculty of Public Health, consistency of exam tasks with education aims for individual subjects is of relevance. Validation of this parameter requires the analysis of content outlines, which is not the subject of this work. Whereas, internal structure validity involves the degree to which psychometric relationships among components within an exam are consistent with the intended meaning of scores for those components [21]. In order to assess internal structure validity, in is necessary to establish cross-correlates between individual subjects. There is, however, a certain critique in the assessment of the measurement using the correlation analysis. It refers to the problem of the measurement scale because school grades are ordinal numbers and thus they do not create an interval scale that is required when calculating a correlation coefficient according to the formula suggested by Pearson. As Stanley Stevens says: "statistical manipulations that empirical data may be subjected to depend on the type of a scale used to order these data" [22]. In educational measurement there is, however, a certain consensus as for the possibilities to use r-Pearson coefficient calculated on the bases of data based on the ordinal scale [23, 24]. Determination of validity in the inter-correlation analysis presented in this work shows that for a vast majority of subjects there are positive dependencies between the results obtained by students in consecutive exams. However, the power of correlation for different pairs of subjects was too great (r-Pearson between 0.09 and 0.44). Results of inter-correlation analysis are to a large extent consistent with the results of assessment consistency of scoring evaluated by the value of Kendall Tau rank correlation coefficient and the measurements of reliability performed using α-Cronbach coefficient. The analysis of validity of educational measurement aims at preventing abuse in interpreting the results of measurement [25]. If a student achieved a high average throughout the course of studies, then the value of such a grade is relevant only in a situation in which it reflects the student's actual achievements in relation to the curriculum requirements in particular. Therefore, one of the important factors when analysing the validity of the measurement is predictive validity which relates to the measurement of consistency of predictive ability of given results to forecast the future of students, e.g. achieving success throughout the course of studies or the graduate's future professional status. For obvious reasons, determining predictive validity in the analysed case is not possible. We do not have detailed data at our disposal that would concern the future of students and graduates in the field of their professional activities after they have completed their studies.

Last but not least feature of a good educational measurement is objectivity which, in its conventional form, means precision with which results of the measurement are a reflection of curriculum requirements, although in a very narrow definition of a problem. Assessing the quality of this criterion brings many analytical problems. Because it is difficult to clearly determine: a) what should be the

leading range of requirements? B) how to establish individual levels of requirements? C) how to normalise and standardise the measuring tools so that when a student obtains a grade it could be assumed that both knowledge and skills were mastered to a sufficient degree? [9]. All of the above questions can be answered practically solely through teachers and lecturers intuitively formulating their requirements, which often are the result of academic tradition or certain personal visions and convictions. They cannot be, however, seen as objective in this situation. Since it is purely a contractual issue establishing the fact that at a given result of a measurement, we may interpret it in a certain way [26]. For instance, the existence of a "satisfactory" grade in a given subject is solely a teacher's assessment of the level of a student's achievements, just as 70% that a student gains is seen as an appropriate solution to the task that represents a given area of science and is established as sufficient to acknowledge it as satisfactory. Colloquial understanding of objectivism as lack of prejudice in the attitude of the examiner is simply a small part of the notion of objectivism of educational measurement [6, 9, 26].

The discussion of individual properties of educational measurement presented above is a well-known model of analysis of any measurement in psychometry. However, the growing significance of educational studies in relation to educating medical personnel and specialists in the field of Public Health requires developing new concepts both in theory and analysis, which would include specificity of this education.

6 CONCLUSIONS

Each of the three analytical strategies – assessing r-Pearson correlation, Tau Kendall and α-Cronbach coefficient, even though having different logical bases and calculation formulae, leads to similar conclusions. A low level of consistency in evaluating students in consecutive years proves that the system of measuring educational outcomes at the faculty of Public Health is insufficient. A critical analysis of currently used solutions allows to plan a detailed evaluation of the currently used criteria of assessment so as to diagnose the weak points in the field of teaching and assessing reliability and validity of the applied methods used to evaluate students' competences at the Faculty of Public Health at MUW.

REFERENCES

[1] van der Vleuten, C.P.M., Dolmans, D.H.J.M., Scherpbier, A.J.J.A. (2000). The need for evidence in education. Medical Teacher 22(3), pp. 246-50.

[2] Harden, R.M., Grant, J., Buckley, G., Hart, I.R. (2000). Best Evidence Medical Education. Advances in Health Sciences Education. Theory and Practice 5(1), pp. 71-90.

[3] Schuwirth, L.W., van der Vleuten, C.P.M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher 33(10), pp. 783-97.

[4] van der Vleuten, C.P.M. (1995). Evidence-based education? Advances in Physiology Education 269(6 Pt 3), p. S3.

[5] Santayana, G. (2011). The Life of Reason: Introduction and Reason in Common Sense: MIT Press.

[6] Niemierko, B. (1999). Pomiar wyników kształcenia. Warszawa: Wydawnictwo Szkolne i Pedagogiczne.

[7] Rowley J. (1996) Measuring quality in higher education. Quality in Higher Education 2(3), pp. 237-55.

[8] Tam, M. (2001). Measuring Quality and Performance in Higher Education. Quality in Higher Education 7(1), pp. 47-54.

[9] Niemierko, B. (2009). Diagnostyka edukacyjna. Warszawa: Wydawnictwo Naukowe PWN.

[10] Kendall, M.G. (1938). A new measure of rank correlation. Biometrika 30(1/2), pp. 81-93.

[11] Babbie, E. (2013). The practice of social research. 13th ed. Belmont: Cengage Learning.

[12] Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika 10(4), pp. 255-82.

[13] Feldt, L.S. (1969). A test of hypothesis that Cronbachs alpha or Kuder-Richardson coefficent 20 is same for 2 tests. Psychometrika 34(3), p. 363.

[14] Niemierko, B. (1975). Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe. 1st ed. Warszawa: Wydawnictwo Szkolne i Pedagogiczne.

[15] Guilford, J.P. (1954). Psychometric methods. 2nd ed. New York: McGraw-Hill.

[16] Goodwin, L.D. (2002). Changing conceptions of measurement validity: an update on the new standards. The Journal of Nursing Education 41(3), pp.100-6.

[17] White, J., Brownell, K., Lemay, J.F., Lockyer, J.M. (2012). "What do they want me to say?" The hidden curriculum at work in the medical school selection process: a qualitative study. BMC Medical Education 12, p. 17.

[18] Ebel, R.L. (1961). Must all tests be valid? American Psychologist 16(10), pp. 640-7.

[19] Reezigt, G.J., Guldemond, H., Creemers, B.P. (1999). Empirical validity for a comprehensive model on educational effectiveness. School Effectiveness and School Improvement 10(2), pp. 193-216.

[20] Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin 52(4), pp. 281.

[21] Meagher, D.G., Pan, T., Wegner, R., Olson, A.T., Overgaard, S.L., Mehle, J.J. (2012). PCAT Reliability and Validity. 3rd ed. San Antonio: Pearson Executive Office.

[22] Stevens, S.S. (1946). On the theory of scales of measurement. Science 103(2684), pp. 677-80.

[23] Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. Psychological Bulletin 87(3), pp. 564-7.

[24] Velleman, P.F., Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. The American Statistician 47(1), pp. 65-72.

[25] Kubielski, W. (2006). Podstawy pomiaru, konstruowania i ewaluacji testu dydaktycznego: Wydawnictwo Wyższej Szkoły Pedagogicznej TWP.

[26] Thorndike, R.L., Angoff, W.H. (1971). Educational measurement: American Council on Education Washington, DC.