ANALYSIS OF RELIABILITY AND VALIDITY OF THE PERFORMANCE ASSESSMENT SYSTEM FOR STUDENTS OF THE SECOND-LEVEL STUDIES IN NURSING - A SEVEN-YEAR RETROSPECTIVE ANALYSIS

M. Panczyk, J. Belowska, A. Zarzeka, H. Rebandel, J. Gotlib

Division of Teaching and Outcomes of Education, Faculty of Health Sciences, Medical University of Warsaw (POLAND)
mariusz.panczyk@wum.edu.pl, jaroslawa.belowska@wum.edu.pl, aleksander.zarzeka@wum.edu.pl, henryk.rebandel@wum.edu.pl, joanna.gotlib@wum.edu.pl

Abstract

Introduction:
Educational measurement at the higher education level is now based on the assessment of a student's progress, with particular emphasis on the measurement of learning outcomes. Evidence-based assessment, as a good practice, should be applied in any well manager higher education institution which attempts to adjust its educational policy to the changing needs and requirements of the labour market.
Aim of study:
Analysis of the reliability and validity of the system of assessing students' performance in subjects terminating with an examination at the second-level studies in the field of nursing.
Materials and Methods:
Examination data of 901 students in the field of nursing who undertook second-level studies at the Faculty of Health Sciences, Medical University of Warsaw (MUW) in the years 2006/07–2012/13 were qualified for the study. Retrospective analysis covered educational performance in ten subjects which terminated with an examination: *Management in nursing, Principles of psychotherapy, Theory of nursing, European nursing, Epidemiology, Medical didactics, Psychiatry, Medical services, Family planning, Law in health care*. The internal cohesion of the assessment system was evaluated with alpha-Cronbach coefficient. The concordance of the educational measurement for individual subjects was evaluated with the use of ANOVA Friedman non-parametric test with Kendall's coefficient of concordance. The theoretical validity assessment was performed with the use of two methods: 1) r-Spearman's rank correlation coefficient to estimate the inter-correlation of educational performance; 2) exploratory factor analysis with Varimax rotation of raw factor loadings.
Results:
The overall reliability measurement for the ten analysed subjects was alpha-Cronbach = 0.605. The total comparative analysis of mean grades obtained by students in the years 2006/07–2012/13 reveals that the worst assessed area was *Epidemiology* (mean grade 3.2 ± 0.73), while the best *Principles of psychotherapy* (mean grade 4.8 ± 0.40). For all the remaining subjects students' mean grades ranged from 3.6 to 4.4 which indicates an average level of concordance (Tau Kendall's coefficient 0.34). Results of the internal structure of students' performance assessment show that two subjects *Principles of psychotherapy* and *Medical services* do not show significant positive inter-correlations with the remaining subjects. The assessment of validity by means of the factor analysis revealed that four subjects form separate domains of the assessed educational performance. The strongest distinctness was registered in the case of *Principles of psychotherapy* while a medium level of distinctness was observed in *Medical services, Psychiatry* and *Theory of nursing*.
Conclusions:
The estimated reliability and validity of the applied methods of nursing students' performance assessment allows to validate and further monitor the quality of the assessment system. In future, it will be necessary to evaluate the criterion validity by estimating the prognostic capacity of the principles of assessment with the use of data concerning further professional careers of graduates.

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

1 INTRODUCTION

The curriculum that is effective at the Medical University of Warsaw (MUW) at the studies of the 2nd degree includes all indications stipulated in the list of standards concerning directional education and regulated by appropriate Regulations of the Minister of Education and Higher Education [1, 2]. Generally, the Regulations state that the studies of the 2nd degree may last no shorter than 4 semesters, the number of classes and training may not be lower than 1300 and the number of ECTS credits (European Credit Transfer and Accumulation System) when divided into the basic and directional content may not be less than 120 [2]. A graduate of the 2nd degree studies has expertise in the field of nursing and other medical sciences. As for skills, a graduate may, for instance: 1) solve professional problems, 2) determine the standards of professional assistance, 3) carry out scientific research into their area of specialisation, 4) organise and supervise nursing care, 5) organise the work of subordinates and their own, according to the current legal regulations, 6) elaborate the assumptions of HR and the plan of employment, 7) elaborate and implement monitoring and assessment tools in the nursing profession, 8) elaborate programmes of health education and realise them with reference to the selected social community [2]. A graduate is prepared for their work in public and private health centres; public local government administration and education – after completing the teaching course (compliant with the standards of teaching preparing for the teacher's position). A graduate has the routine of constant self-development and improvement in their profession, and is prepared to undertake studies of the 3rd degree (PhD) [1].

Curriculum of studies of the 2nd degree covers 27 subjects and additional practices. Among the social subjects, 5 of them end with an exam, whereas in the area of specialized care – there are two. Additionally, on the list of non-standard subjects that ended in an exam, there were between two to four, depending on the year [1]. Studies of the 2nd degree at the Nursing department are completed with a diploma exam that should include checking knowledge and practical skills acquired throughout the course of studies [2]. A detailed list of exam subjects included in the curriculum of full-time studies of the 2nd degree at the Nursing department at MUW between the years 2006/7 – 2012/13 are presented in Table 1. A summary of individual exam subjects for individual years are collected in Table 2.

Assessing students is seen as one of the more important elements of the whole system of education. On the one hand, it determines the degree to which a student achieved the assumed level of learning, on the other, however, it can also become a tool measuring the quality of the process of education [3]. Regardless of the purpose any evaluation serves, it always is connected with more or less systematic observation data collection that is to lead to some conclusions concerning the features and properties of the evaluated student [3]. For this process to be a highly objective source of information on learning outcomes, it must meet certain requirements specified as the features of educational diagnosis. And reliability and validity are listed among some of such features that allow to optimize educational measurement [4-6].

Table 1. A list of subjects ending in an exam and included in the curriculum of full-time studies of the 2nd degree at the Nursing department at the Medical University of Warsaw between the years 2006/07 – 2012/13.

Year of studies	Subjects terminating with an examination	Content	ECTS credits
I	Management in nursing	Social sciences	6
	Principles of psychotherapy	Social sciences	3
	Theory of nursing	Social sciences	4
	European nursing	Social sciences	3
	Epidemiology	Non-standard	3
II	Medical didactics	Social sciences	9
	Psychiatry	Specialized care	5
	Medical services	Non-standard	2
	Family planning	Non-standard	2
	Law in health care	Non-standard	2
	Oncology	Specialized care	5

Table 2. Individual subjects terminating in an exam included in the curriculum for consecutive years beginning their full-time studies at the Nursing department at the Medical University of Warsaw.

Subjects terminating with an examination	year	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13
Management in nursing*	I	X	X	X	X	X	X	X
Principles of psychotherapy*		X	X	X	X	X	X	X
Theory of nursing*		X	X	X	X	X	X	X
European nursing*		X	X	X	X	X	X	X
Epidemiology*		X	X	X	X	X	X	X
Medical didactics*	II	X	X	X	X	X	X	X
Psychiatry*		X	X	X	X	X	X	X
Medical services					X	X	X	
Family planning					X	X	X	
Law in health care					X	X	X	X
Oncology								X

** subjects included in the analysis and presented in this scientific description*

2 AIM OF STUDY

Analysis of reliability and validity of the system of assessing students' performance in subjects terminating with an examination at the second-level studies in the field of nursing at MUW.

3 MATERIALS AND METHODS

Overall, there were 901 students who began their studies at the Nursing department of the 2nd degree between the years 2006/07 and 2012/13. Total coefficient of attrition during the course of studies was 26.1%. The age average was between 26.3 ± 7.00 years. A detailed characteristic of the population of students who study Nursing at MUW between the academic years of 2006/07 and 2012/13 is presented in Table 3. The analysis included seven basic subjects that were assessed: *Management in nursing, Principles of psychotherapy, Theory of nursing, European nursing, Epidemiology, Medical didactics,* and *Psychiatry.*

Table 3. Characteristic of students who began their studies at the Nursing department of the 2nd degree at the Medical University of Warsaw between 2006/07 – 2012/13.

Year of beginning studies	Mean age	SD	*N* students	*N* women	*N* men
2006/07	27.1	7.00	89	84	5
2007/08	26.1	6.98	146	137	9
2008/09	27.1	7.17	167	152	15
2009/10	24.4	5.19	155	149	6
2010/11	27.0	7.37	122	117	5
2011/12	25.9	7.33	102	94	8
2012/13	26.2	7.93	120	114	6
TOTAL	**26.3**	**7.00**	**901**	**847**	**54**

In order to evaluate reliability of the system of students' assessment, a method of internal compliance was used that was suggested by Cronbach (Kruder-Richardson coefficient for a test comprising of two sub-category positions) [7]. Based on Nunnally's criterion, the assumed level of reliability measured by α-Cronbach coefficient, should be at least 0.70 [8]. So as to assess the inner-scale compliance for individual subjects, the matrix of inter-correlation was established and, as a

criterion of satisfactory level of assessment coherence, an average value of r–Spearman greater than 0.20 was assumed [9]. Whereas is order to assess the compliance of assessment of individual students for consecutive exam subjects, a non-parametric ANOVA Friedman test was used to compare the dependent variables with Kendall coefficient of compliance.

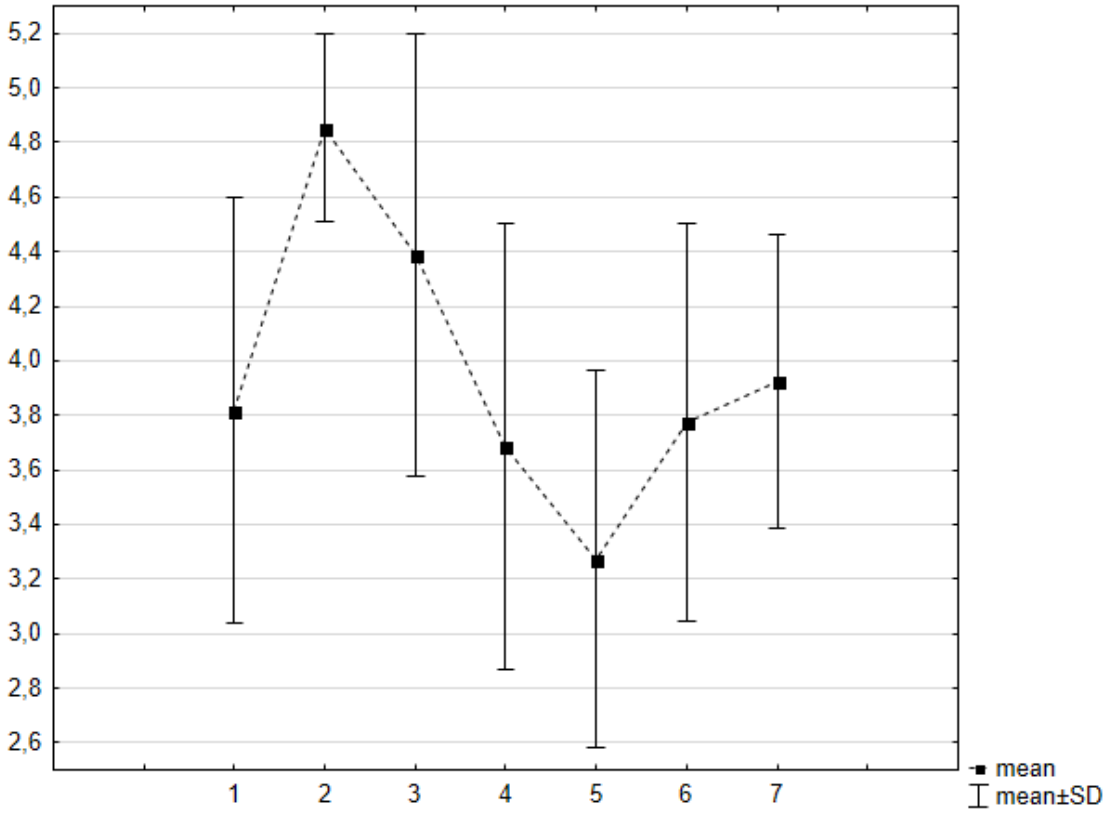
So as to assess validity of the applied measuring scale of students' achievements, a method of indirect assessment of theoretical validity was used [10, 11]. So as to single out the mutually correlated positions on the scale, exploratory factor analysis was used [12]. In order to single out a number of factors, Kaiser's criterion was applied and the scree criterion of Cattell [13, 14]. To determine which subjects create a mutually correlated grouping for the singled out factors, a method of rotation of crude factors of loading *Varimax* was used [13, 15].

For calculations, STATISTICA package was used, version 12 (StatSoft, Inc.) according to the licence for MUW. For all analyses, the level of relevance was assumed *a priori* as $\alpha = 0.05$.

4 RESULTS

The analysis of homogeneity in assessing students shows that there is a different level of difficulty in individual subjects. On the one hand, the results of the statistics presented in Figure 1 show that, particularly in case of one subject, a visibly lower average of exam grades was observed in comparison with other subjects (average grade for *Epidemiology* was 3.30). On the other hand, it was discovered that for *Principles of Psychotherapy*, the average grade reached the value of 4.90. The results of comparative analysis using a non-parametric ANOVA Friedman test ($chi^2 = 2222.545$; $P < 0.000001$) show a total compliance of measurement on the level of 0.41. A detailed summary of statistical parametres for the results of assessing subjects ending in an exam is presented in Table 4.

Figure 1. Structure of students' grades for subjects ending in an exam and included in the curriculum of full-time studies at the Nursing department at the Medical University of Warsaw between the years 2006/07 – 2012/13.



1 – Management in nursing, 2 – Principles of psychotherapy, 3 – Theory of nursing, 4 – European nursing, 5 – Epidemiology, 6 – Medical didactics, 7 – Psychiatry

Table 4. Selected parametres of descriptive statistics concerning evaluating subjects ending in an exam at the Nursing department of full-time studies at the Medical University of Warsaw between the years 2006/07 – 2012/13.

Subject	Mean	SD	CV (%)
Management in nursing	3.8	0.78	20.5
Principles of psychotherapy	4.9	0.35	7.1
Theory of nursing	4.4	0.81	18.5
European nursing	3.7	0.82	22.1
Epidemiology	3.3	0.69	21.2
Medical didactics	3.8	0.73	19.3
Psychiatry	3.9	0.54	13.7

SD – standard deviation; CV – coefficient of variation

The evaluated inner compliance of the measurement calculated using Cronbach formula, shows that the level of homogeneity of the assessing system is lower than assumed. Alpha reliability coefficient was 0.559 and thus surpassed the threshold of 0.70. However, in case of one subject – *Principles of psychotherapy*, a relatively low discrimination index was noted, measured using discriminating power. For all other subjects, this index oscillated around the values of 0.20-0.40. Moreover, it was also noted that the assessment results in *Epidemiology* have the highest influence on the inner compliance of educational measurement. A detailed summary of the results of reliability analysis and the positions is presented in Table 5.

Table 5. Results of reliability analysis and the positions on the measurement scale of all achievements of the students for subjects ending in exams at the Nursing department at MUW between the years 2006/07 – 2012/13

Subject	Discriminating power	Reliability after removal **
Management in nursing	0.351	0.492
Principles of psychotherapy	-0.001*	0.586
Theory of nursing	0.230	0.547
European nursing	0.383	0.476
Epidemiology	0.405	0.472
Medical didactics	0.338	0.498
Psychiatry	0.210	0.545

** position lowering the reliability of the scale measuring the achievements of the students*
*** alpha-Cronbach coefficient*

The results of inter-correlation analysis show that in individual areas of studies, positive dependencies between the learning outcomes of the students are noted for only some of the studied subjects. In case of *Principles of psychotherapy*, as many as three cases of lack of correlation with the grades achieved by students in other exam subjects were noted. Moreover, a significant negative correlation was observed between the exam results in this subject and *Theory of nursing* (r_s = –0.17). The highest mean value of correlation was shown for *Epidemiology* (mean r_s = 0.20) and *European nursing* (mean r_s = 0,19). The results of the analysis of correlation thus confirm the findings discovered while estimating reliability using α-Cronbach coefficient. A detailed summary of the results of this analysis is presented in Table 6.

On the basis of the obtained scree diagram, according to the Cattell criterion, a structure comprising of two factors was assumed as the optimum in the factor analysis. Charge values obtained using *Varimax* rotation allowed to determine individual components of factors 1 and 2. The following subjects were singled out as individual concentration of mutually correlated grades: *Epidemiology, Medical didactics, Management in nursing* and *European nursing*. Moreover, results in *Principles of psychotherapy* should be considered a separate factor. Other subjects cannot be clearly assigned to any of the factors (Table 7). Also, according to the diagram presented in Figure 2, the results of assessing students in *Principles of psychotherapy* should be assumed as significantly different.

Table 6. Results of the analysis of inter-correlation for the grades of students in individual subjects ending in an exam at the Nursing department of the 2nd degree at MUW between the years 2006/07 – 2012/13.

	1	2	3	4	5	6	7	Mean correlation
1. Management in nursing	–	0.05*	0.11	0.28	0.26	0.22	0.13	0.18
2. Principles of psychotherapy	0.05*	–	-0.17	0.10	0.01*	-0.04*	0.07	0.00
3. Theory of nursing	0.11	-0.17	–	0.12	0.26	0.16	0.16	0.11
4. European nursing	0.28	0.10	0.12	–	0.26	0.32	0.07	0.19
5. Epidemiology	0.26	0.01*	0.26	0.26	–	0.21	0.17	0.20
6. Medical didactics	0.22	-0.04*	0.16	0.32	0.21	–	0.09	0.16
7. Psychiatry	0.13	0.07	0.16	0.07	0.17	0.09	–	0.11

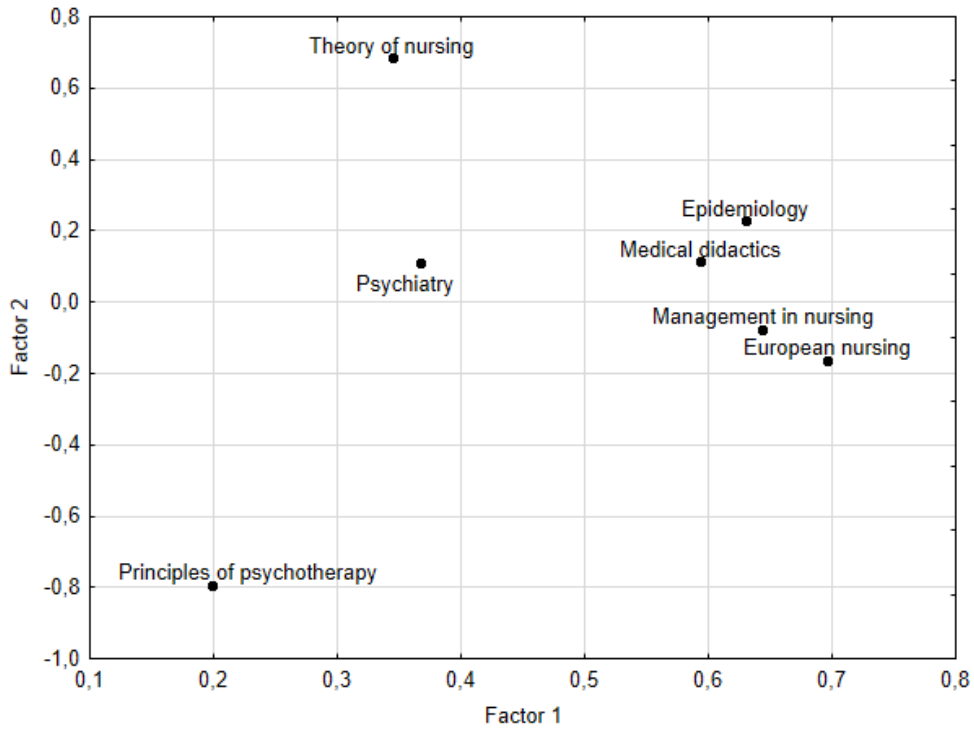
** statistically insignificant (P > 0.05)*

Table 7. Factor structure of the position of measurement scale of learning outcomes for subjects ending in an exam, created on the basis of assessing the factor charges from *Varimax* rotation.

Subject	Factor 1	Factor 2
Management in nursing	0.643404*	0.075209
Principles of psychotherapy	0.198448	0.794345**
Theory of nursing	0.344615	-0.683852
European nursing	0.695788*	0.161993
Epidemiology	0.630516*	-0.229759
Medical didactics	0.593418*	-0.116075
Psychiatry	0.366572	-0.108449

** a subject forming factor 1 ** subjects forming factor 2*

Figure 2. Scatterplot of individual factor charges for subjects ending in an exam obtained in the basis of *Varimex* rotation



5 DISCUSSION

Measurement impartiality, i.e. independence of the measured situation means providing all students with the same (fair) conditions while assessing their achievements. Equal treatment of all students in consecutive years means independent evaluation of their achievements, regardless of their previous score in previous education cycles, their previous schools / universities or student groups they studied in. An important element of measurement impartiality is creating appropriate conditions during the exam and selecting such methods of evaluation of students' achievements that would ensure a comparable degree of independence of the measured situation in consecutive years [16-18].

Validity of scoring (objectivism of scoring) is directly connected with the measurement impartiality, as it is understood as adequacy of a given measuring scale towards the assessed properties. In practice, achieving a high accuracy of scoring on the one hand depends on the manner in which exam tasks are prepared and the quality of rubrics of evaluating a described key (constructive causes), and on the other hand, it is a derivative of competences, professional experience and personality of the examiner (personal causes) [16]. The most frequent source of inconsistency related to the results of measurement is excessive austerity / leniency of the examiner as well as their propensity to average [4]. As can be seen in the analysis of grades in individual subjects, in case of *Epidemiology*, students achieved significantly lower grades in comparison with other subjects (a harsh examiner), while in *Principles of psychotherapy* grades were extremely high (an examiner that was too lenient). At the same time, in case of *Epidemiology*, a wide array of inter-changeability of results was observed, which naturally reflects the differences of achievements in the studied group of students. Extreme values do not have to be the evidence of the lack of objectivism in scoring, but they may result from the fact that a given subject requires complex and difficult competences that must be mastered (as is the case in *Epidemiology* where there was a very low average of grades in students) or on the contrary – achieving any learning outcomes is relatively easy for students in this case (high average of grades, as is the case in *Principles of psychotherapy*). The differentiated level of expectations towards students in individual subjects may, to a certain degree, justify the result of analysis of students' achievements, pointing to certain statistically relevant deviation in the consistency of assessing. However, compliance of a grade within the same subject was different in consecutive years. It should be expected that regardless of the year, a good student – as opposed to the less talented one – will have significantly higher results in all subjects. The measure of this compliance is Tau Kendall coefficient that is based on the difference between probability of the fact that two variables are arranged in the same order within the observed data and the probability that their arrangement differs (the value of this coefficient was 0.41) [19]. Error in the range of accuracy in scoring may be connected with excessive rigidity of evaluation criteria in a situation when score is introduced not for educational reasons but it merely results from the adequacy of a solution to the task, which does not necessarily falls within the knowledge and skills provided for the measured learning outcome. This problem will be of particular importance while assessing reliability and validity of the measurement [6].

Reliability of the measurement means repeatability of the obtained results under certain conditions. In order to assess whether a given measurement is reliable, various analytical methods might be applied. A degree of correlation between results of individual exams of their part that is most frequently used practice (e.g. *odd-even* or *split-half reliability*) [20], or it may be the assessment of internal consistency of the results of the measurement for all exams (alpha-Cronbach) [7]. As can be seen from the presented results of studies into reliability using alpha-Cronbach, system of students' achievements was characterised by insufficient level of reliability (non-fulfilled Jum Nunnaly's criterion $\alpha > 0.70$, although some scientists also accept the value of 0.600 [8]). It needs to be emphasised that inappropriate selection of exam methods that are the basis of assessment, and inappropriate construction of testing tasks in particular, contribute to the decrease in the measurement of reliability. A student may not have enough possibilities to present their achievements in a given field if an exam considerably narrows the content which are intended to evaluate learning outcomes. Log reliability of the evaluating system may take place particularly in such cases where there is an extremely high degree of differentiation of the assessed effects, which in most cases may concern, especially in general education, rather remote features and properties of the test takers. Another reason of low value of α coefficient (< 0.70) may be a high number of random errors in the results of a given measurement. Random fluctuation of learning outcomes described in classic theory of an exam test may lower the value of α coefficient [21]. With α value < 0.7, random errors account for more than 30% of the variation of the obtained results and any measurement applied in such conditions may, according to Guilford, only be used while applying inter-group comparison and not in individual differentiation [22]. Moreover, it should be noted that aiming solely at achieving high values of α coefficient does not solve the problem

of reliability because a high α value only means minimization of the influence of random errors on the obtained results, yet it does not provide certainty as far as the existence of systematic errors, sometimes very serious ones, which are connected with the partiality of the measurement [22].

The results of the inter-correlation analysis show that not for all subjects there are positive dependencies between the results achieved by students in their exams. Moreover, the power of correlation for individual pairs of subjects was varied and a mean value of r-Spearman correlation coefficient oscillated between 0.11-0.20, and was significantly lower only in cases of results in *Principles of psychotherapy*. Also, that was the only subject for which low and insignificant correlations with other learning outcomes were noted. These findings thus confirm previous observations concerning the analysis of reliability that pointed to insufficient discriminatory power of grades in *Principles of psychotherapy*. Improvement in cohesion parametres of evaluation requires a detailed quality analysis of the already applied methods of educational measurement for all exam subjects.

Apart from establishing the degree of reliability of the measurement that refers to the question "*How to measure*?", in creating good evaluating tools it is important to determine the validity of measurement that would answer the question "*What is measured*?". Validity in this field should be understood as a degree of compliance with which a measuring tool measures what it has been designed to measure [23]. Thus we may talk about the usefulness of a given method in evaluating a certain set of properties and features of a test taker [24]. If a selected method really checks the skills of a student's ability to adjust to a selected tool ("*What do they want me to say*?"), then in such a case evaluation is not directed at these properties that we wish to measure [25]. No precise method of measuring validity exists, there is only its indirect evaluation, which is usually based on the application of one of the three concepts thanks to which it is possible to determine the validity of measurement: content-based, theoretical and criteria-based [4, 6, 21]. In case of exams included in the curriculum at the Nursing department, compliance of test tasks with the aims of learning for individual subjects is important as far as content (curriculum) validity is concerned. Validation of this parametre thus requires the analysis of the curriculum, which is not the subject of this study. However, the analysis of theoretical validity concerns the degree to which individual components, i.e. parts that are included in the whole of the measurement, are mutually correlated [26]. The results of the factor analysis are to a great extent coherent with the results of the compliance evaluation of scoring and the measurement of reliability using alpha-Cronbach coefficient, which showed that grades obtained in the subject of *Principles in psychotherapy* are a different measurement of features and properties of test takers when contrasted with the results of exams in other subjects.

The aim of the analysis of validity of the educational measurement is to prevent abuse in interpretation of the results of measurement [27]. If a student achieved a high average throughout the course of study, then the value of such assessment is relevant only if it actually reflects their achievements, in particular with reference to the curriculum. That is why, one of the most important aspects of the analysis of the quality of measurement is to assess the predictive validity that refers to evaluation of the degree to which certain learning outcomes may serve as predictors of the students' future fate, e.g. their professional status as graduates. For obvious reasons, determining predictive validity in the analysed case is not fully possible since we do not have detailed data concerning the future of graduates as for their professional activities following their course of studies at our disposal. Other valuable guidelines as for validity of the evaluation system may also be obtained in the process of the analysis of diagnostic properties of educational measurement. In order to do so, it is necessary to check the competences achieved by the student / graduate in the test that would be carried out externally. If, in the future, for all graduates of the direction, the obligation to pass the National Nursing Exam (an equivalent of the Medical Final Exam) is introduced should they wish to receive the right to perform their profession, then it will be possible to verify the quality of the didactic process and diagnostic validity of the system of students' assessment at a given university. The analyses of the quality of education of nurses in the USA may serve as an example here; there, the external criterion of evaluation of professional preparation is passing the National Council Licensure Examination for Registered Nurses (NCLEX-RN) by the graduate [28. 29].

The results of the analysis of validity raise certain reservations. In some cases it can be assumed that the results achieved by students do not reflect the actual features and properties of the test takers (see *Principles of psychotherapy*). It needs to be remembered, however, that in the deepened analysis of the curriculum, a critical evaluation should be performed as for how far such a subject really represents the range of competences relevant to future clinical education. A situation in which rather good reliability accompanies insufficient validity results in lowering credibility of the whole system of competence evaluation. This is of particular importance when evaluating those students whose average

score throughout their course of studies places them near the lower border of the scale (grade average ≈ 3.0). Due to the fact that these are the weakest students, there is a high probability of insufficient level of competences represented by this group of graduates of the Nursing department.

6 CONCLUSIONS

The estimated reliability and validity of the applied methods of nursing students' performance assessment allows to validate and further monitor the quality of the assessment system. In future, it will be necessary to evaluate the criterion validity by estimating the prognostic capacity of the principles of assessment with the use of data concerning further professional careers of graduates.

REFERENCES

[1] Przewodnik dydaktyczny dla studentów kierunku pielęgniarstwo studia II stopnia. 2008, Warszawa: Warszawski Uniwersytet Medyczny.

[2] Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego z dnia 9 maja 2012 r. w sprawie standardów kształcenia dla kierunków studiów: lekarskiego, lekarsko-dentystycznego, farmacji, pielęgniarstwa i położnictwa (Dz.U. 2012 nr 0 poz. 631).

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

[4] Niemierko, B. Diagnostyka edukacyjna. 2009, Warszawa: Wydawnictwo Naukowe PWN.

[5] Norman, G.R., Vleuten, C., Newble, D.I. International handbook of research in medical education. Vol. 7. 2002: Springer.

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

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

[8] Nunnally, J.C., Bernstein I.H. Psychometric theory. 3 ed. Vol. 226. 1967, New York: McGraw-Hill.

[9] Jankowski, K., Zajenkowski, M. Metody szacowania rzetelności pomiaru testem [in] Psychometria - podstawowe zagadnienia, K. Fronczyk, Editor. 2009, Vizja Press & IT: Warszawa, pp. 84-110.

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

[11] Tarnowski, A., Fronczyk, K. Szacowanie trafności [in] Psychometria - podstawowe zagadnienia, K. Fronczyk, Editor. 2009, Vizja Press & IT: Warszawa, pp. 140-160.

[12] Zakrzewska, M. Analiza czynnikowa w budowaniu i sprawdzaniu modeli psychologicznych. 1994, Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu.

[13] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika 23(3), pp. 187-200.

[14] Fabrigar, L.R., et al. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods 4(3), pp. 272.

[15] Stanisz, A. Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach medycznych: Tom 3. Analiza wielowymiarowa. 2007, Kraków: StatSoft Polska.

[16] Rowley, J. (1996). Measuring quality in higher education. Quality in Higher Education 2(3), pp. 237-255.

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

[18] Sood, R., Singh T. (2012). Assessment in medical education: Evolving perspectives and contemporary trends. National Medical Journal of India 25(6), pp. 357-364.

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

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

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

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

[23] Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research 62(3), pp. 229-258.

[24] 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.

[25] White, J., et al. (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: pp. 17.

[26] Meagher, D.G., et al. PCAT Reliability and Validity. 3rd ed. 2012, San Antonio: Pearson Executive Office.

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

[28] Seldomridge, L.A., DiBartolo M.C. (2004). Can success and failure be predicted for baccalaureate graduates on the computerized NCLEX-RN? Journal of Professional Nursing 20(6), pp. 361-368.

[29] Crow, C.S., et al. (2004). Requirements and interventions used by BSN programs to promote and predict NCLEX-RN success: A national study. Journal of Professional Nursing 20(3), pp. 174-186.