ANALYSIS OF RELIABILITY AND VALIDITY OF THE PERFORMANCE ASSESSMENT SYSTEM FOR BACHELOR'S DEGREE STUDENTS IN MIDWIFERY: A SINGLE-CENTRE STUDY

ANALIZA RZETELNOŚCI I TRAFNOŚCI SYSTEMU OCENY OSIĄGNIĘĆ STUDENTÓW POŁOŻNICTWA NA STUDIACH PIERWSZEGO STOPNIA: BADANIE JEDNOOŚRODKOWE

Mariusz Panczyk, Jarosława Belowska, Aleksander Zarzeka, Łukasz Samoliński, Joanna Gotlib

Division of Teaching and Outcomes of Education
Medical University of Warsaw, Poland

DOI: https://doi.org/10.20883/pielpol.2016.49

ABSTRACT

Introduction. The system of assessment of students' performance needs to comprise certain psychometric characteristics, including in particular reliability and validity, so that it can be a highly objective source of knowledge of students' learning outcomes.
Aim. Analysis of reliability and validity of the performance assessment system for Midwifery students who started a Bachelor's degree programme at Medical University of Warsaw between 2005-06 and 2012-13.
Material and methods. A retrospective study enrolling a group of 922 students of eight subsequent full education cycles. The authors collected detailed data on grades for twenty courses that ended with an exam throughout the course of studies, divided into four groups according to the criteria specified in the education standards. Reliability (Cronbach's alpha coefficient), as well as theoretical (factor analysis) and criterion validity (the Pearson correlation matrix) were assessed. IBM® SPSS® Statistics version 23 was used for calculation.
Results. Cronbach's alpha coefficient for each year exceeded the assumed threshold of 0.700. Total reliability of the assessment of students' performance for the period considered amounted to 0.805. The factor analysis of students' grades for 20 examination courses demonstrated a five-factor structure, which is far from the assumptions resulting from the education standards (four groups of effects). The highest level of criterion validity was observed for the D group courses ("Education in Specialist Care"), with average values of Pearson's correlation coefficient (r) amounting to 0.20.
Conclusions. A good level of reliability accompanied by a low level of validity leads to a decrease in credibility of the entire system of assessment of Midwifery students' competencies. This may, in some cases, increase the risk that there would be persons with insufficient level of initial competencies among Midwifery graduates.

KEYWORDS: midwifery, educational measurement, graduate education, reproducibility of results, professional competence.

STRESZCZENIE

Wprowadzenie. Aby system oceniania studentów był wysoce obiektywnym źródłem informacji na temat osiągniętych efektów kształcenia musi on posiadać pewne właściwości psychometryczne, do których należą przede wszystkim rzetelność i trafność.
Cel. Analiza rzetelności i trafności systemu oceny osiągnięć studentów położnictwa, którzy podjęli kształcenie na studiach pierwszego stopnia na Warszawskim Uniwersytecie Medycznym w latach 2005/06 – 2012/13.
Materiał i metody. Badanie retrospektywne obejmujące grupę 922 studentów z ośmiu kolejnych, pełnych cykli kształcenia. Zebrano szczegółowe wyniki dotyczące uzyskanych ocen z dwudziestu przedmiotów kończących się egzaminem w całym toku studiów w podziale na cztery grupy zgodnie z kryteriami zapisanymi w standardach kształcenia. Oceniono rzetelność (współczynnik alfa-Cronbacha), trafność teoretyczną (analiza czynnikowa) i kryterialną (macierz korelacji Pearsona). Obliczenia z użyciem programu IBM® SPSS® Statistics wersja 23.
Wyniki. Dla każdego rocznika osiągnięto wartość współczynnika alfa-Cronbacha większą niż zakładany próg 0,700. Łączna zgodność oceny osiągnięć studentów dla całego analizowanego okresu wyniosła 0,805. Wyniki analizy czynnikowej ocen studentów z 20 przedmiotów egzaminacyjnych wskazują na pięcioczynnikową strukturę, co odbiega od założeń wynikających ze standardów kształcenia (cztery grupy efektów). Najwyższy poziom trafności kryterialnej zaobserwowano dla przedmiotów z grupy D („Nauki w zakresie opieki specjalistycznej"), dla której średnia wartość współczynnika r-Pearsona wyniosła 0,20.
Wnioski. Dobry poziom rzetelności z równoczesną niską trafnością skutkuje obniżeniem wiarygodności całego systemu oceniania kompetencji studentów położnictwa. Może to w niektórych przypadkach zwiększać ryzyko obecności w grupie absolwentów kierunku osób o niedostatecznym poziomie kompetencji wyjściowych.

SŁOWA KLUCZOWE: położnictwo, ocena wiadomości, szkolnictwo wyższe, powtarzalność wyników, kwalifikacje zawodowe.

Introduction

A Midwifery Curriculum for Bachelor students at Warsaw Medical University (WMU) covers all principles defined in standards relating to the major studies and regulated by the applicable Regulation of the Minister of Science and Higher Education [1]. General requirements of the Regulation say that Bachelor's degree programmes last at least six semesters, the number of class hours and hours of practice amounts to at least 4720 hours and there are at least 180 ECTS credits (*European Credit Transfer and Accumulation System*) broken down into basic and specialised contents [1]. The programme has a practical profile and the major in Midwifery is a part of education in medical, health, and sports sciences [2]. Bachelor's degree graduates have specialist knowledge of midwifery and other medical sciences. They have the following skills: (1) providing health services to pregnant women, women in labour, post-partum women, and new-born infants, among others; (2) recognising pregnancy, taking care of pregnant women and monitoring pregnancy; (3) taking the necessary measures in urgent situations until a doctor arrives; (4) taking care of a mother and a new-born infant, monitoring the postnatal period and examining a new-born infant; (5) cooperating with the medical personnel; (6) carrying out educational and health activities such as preparation for family life, family planning methods, protection of motherhood and fatherhood, preparation for parenthood and childbirth [1, 2].

The Bachelor's degree curriculum includes a total of 40 courses (2420 hours), 20 of which end up with a final test equivalent to an exam. The curriculum also includes a compulsory practical training (1100 hours) and internship (1200 hours). According to the education standards, all teaching outcomes were divided into four categories (A, B, C, and D) [1]. See **Table 1** for a detailed list of courses included into the curriculum of the full-time Bachelor's degree programme in Midwifery at WMU between 2005–06 and 2012–13.

Table 1. List of courses contained in the standards of education for Bachelor's degree programme in Midwifery during the period 2005/06 - 2012/13 at the Medical University of Warsaw

Learning outcomes group	Course	Exam
A. Basic sciences	Anatomy	•
	Physiology	•
	Pathology	
	Embryology and Genetics	
	Biochemistry and Biophysics	
	Microbiology	•
	Parasitology	•
	Pharmacology	•
	Radiology	
B. Social Sciences	Psychology	•
	Sociology	
	Pedagogics	•
	Law	
	Public health	•
	Philosophy and Ethics in Midwifery	
	Foreign language	•
C. Sciences in the basics of maternity care	Basics of maternity care	•
	Health promotion	•
	Primary health care	•
	Dietetics	
	Physical examination	
	Research in obstetrics	
	Optional courses to choose from: nosocomial infections, sign language and the promotion of mental health	
D. Science in the field of specialist care	Obstetrical techniques and care during childbirth	•
	Obstetrics and Maternity care	•
	Gynecology and Gynecological care	•
	Neonatology and Neonatal Care	•
	Paediatrics and Paediatric Nursing	•
	Internal medicine	•
	Surgery	
	Psychiatry	•
	Anesthesiology and life-threatening situations	
	Rehabilitation in obstetrics, gynecology and neonatology	•
	Basics of medical emergency	

Source: author's own analysis

Assessment of students constitutes one of the most important elements of the entire system of education. On the one hand, it defines the degree to which students achieve the expected learning outcomes and on the other hand, it may also measure the quality of the education process [3]. Regardless of the purpose of the assessment, it is always, more or less, associated with a systematic collection of observational data leading to conclusions on the features and characteristics of a particular student [3]. In order to make the process a highly effective source of information on achieved learning outcomes, it has to meet certain criteria referred to as features of the educational diagnosis. Reliability and validity constitute the core features of educational measurement, allowing for its evaluation and optimisation [4–6].

Aim

Analysis of reliability and validity of the performance assessment system for Midwifery students who started a Bachelor's degree programme at Warsaw Medical University between 2005–07 and 2012–13.

Material and methods

The retrospective study involved data on the course of studies comprising a group of 922 students (women

constituted 100% of all) who started studying midwifery in a full-time Bachelor's degree programme at the Faculty of Health Science, WMU, between the academic years 2005–06 and 2012–13 (eight full training cycles). Among the study group, the total failure rate made up 16.7% and the number of delayed graduation accounted for 7.9%.

For each student, the authors collected detailed data on grades for twenty courses that ended with an exam throughout the course of studies, divided into four groups according to the criteria specified in the education standards. The data were collected from the Central Database of Students whose aim is to support administration handling of students and course of studies.

In line with the position of the Ethical Review Board, WMU, the approval of the Board is not necessary to conduct retrospective studies, surveys, and other non-invasive activities.* The present authors obtained the consent of the Local Controller of the Personal Data for processing of personal data of WMU students.

An analysis of reliability, validity, and one-dimensionality of particular groups of courses were used to assess psychometric characteristics of the performance assessment system for Bachelor's degree students of Midwifery. An analysis of internal consistency of students' achievements with the use of Cronbach's formula was used to assess reliability (see [7]). In compliance with Nunnally's criterion, Cronbach's alpha coefficient was established at $\alpha > 0{,}70$ [8]. Furthermore, a value of the discrimination index was established for each course to assess how grades for a particular course influence the total consistency of the measurement of students' performance. A threshold value for this index was established at 0.20 [9].

Validity of the students' performance assessment system was evaluated with the use of two different analytical approaches:

- Estimation of theoretical validity (also called internal validity) using exploratory factor analysis. It was assessed whether a factor structure comprised four parts, which would reflect the number of selected groups of courses. Meeting of the assumptions of factor analysis was checked, Bartlett's test of sphericity was performed, degree of variance homogeneity was estimated, correlation matrix determinant and KMO measure of sampling adequacy (Kaiser-Meyer-Olkin index) were established.
- Estimation of criterion validity which was based on the degree of correlation between selected courses and groups of courses. Values of Pearson's correlation coefficients (r) were established to verify the assumption mentioned above.

The analysis of one-dimensionality of each group of courses was conducted with principal component analysis. It was assumed that a group of courses is one-dimensional if it meets Kaiser criterion (established eigenvalues exceed the value of 1 only once) and reproducibility of variability of indicator variables with the first principle component exceeds 40% [10].

All statistical calculations were carried out using IBM® SPSS® Statistics version 23. The significance level for each analysis was established *a priori* at $\alpha = 0.05$.

Results

For each of the seven years, Cronbach's alpha coefficient exceeded the determined threshold of 0.700, with only one case of alpha less than 0.800 (year of 2007-08, alpha = 0.790). However, the total consistency of assessment of students' achievements for the entire period amounted to 0.805 after standardisation. The analysis of measurement of students' performance with regard to changes in the value of alpha coefficient after the elimination of particular courses showed that removal of two of them (*Psychiatry* and *Rehabilitation in obstetrics and gynaecology*) may slightly increase the value of this coefficient. See **Table 2** for details on the results of reliability analysis.

Table 2. Analysis results of reliability and discrimination power for courses culminating in an examination during the Bachelor's degree programme in Midwifery

Course	Learning outcomes group	Discrimination power	Cronbach's alpha when removed	Cronbach's alpha
Anatomy	A	0.410	0.795	0.532
Physiology		0.433	0.794	
Microbiology		0.439	0.794	
Parasitology		0.359	0.798	
Pharmacology		0.261	0.804	
Psychology	B	0.401	0.796	0.438
Pedagogics		0.463	0.792	
Public health		0.216	0.804	
Foreign language		0.386	0.796	
Basics of maternity care	C	0,444	0,793	0,397
Health promotion		0.304	0.801	
Primary health care		0,363	0,799	

* Detailed information and model documents of the Ethical Review Board of Warsaw Medical University are available at: https://komisja-bioetyczna.wum.edu.pl/content/szczegółowe-informacje-oraz-wzory-dokumentów (date of access: November 27, 2015).

Obstetrical techniques and care during childbirth	D	0,525	0,789	0,669
Obstetrics and Maternity care		0,562	0,786	
Gynecology and Gynecological care		0,348	0.799	
Neonatology and Neonatal Care		0,427	0,794	
Paediatrics and Paediatric Nursing		0.358	0.798	
Internal medicine		0.409	0.795	
Psychiatry		0.193*	0.808	
Rehabilitation in obstetrics, gynecology and neonatology		0.143*	0.807	

* below the minimum required

Source: author's own analysis

Meeting the assumptions of the method was checked before theoretical validity was estimated with factor analysis. Null standard deviation was not observed for any course and homoscedasticity was confirmed (Levene's test for homogeneity of variances, $p > 0.05$). The value of correlation matrix determinant was close to zero (0.023). Moreover, the criterion of sphericity was met since it was found that the correlation coefficient matrix was not an identity matrix (Bartlett's test, $p < 0.0001$). The last criterion for factor analysis was checked with the use of Kaiser-Mayer-Olkin test assessing the anticipated reduction of measurement scale. KMO index of sampling adequacy amounted to 0.839, meeting the assumptions for this criterion (KMO > 0.500).

In the exploratory factor analysis, students' grades for 16 examination courses were divided into five factors according to Kaiser criterion, which was not consistent with the division into four groups. It was found that variables grouped into five factors explained a total of 50.0% of the entire variance (**Table 3**).

Table 3. Participation of explained variance for each factor – the five-factor solution

Factor	Eigenvalue	Participation of explained variance (%)
1	4,490	22,449
2	1,687	8,435
3	1,439	7,194
4	1,260	6,301
5	1,130	5,651
Total	----	50,029

Source: author's own analysis

Varimax orthogonal rotation of raw factor loadings was carried out to facilitate interpretation of the obtained solution. Component 1 comprised 13 out of 20 courses and no clear-cut factor solution was found for another four courses. Moreover, further three courses fell beyond Component 1. Particular courses did not form structurally separate components that would be consistent with the theoretically assumed division resulting from the standards of education. See **Table 4** for a detailed breakdown of the results of factor analysis with *Varimax* rotation of loadings.

Table 4. The rotation matrix using the *Varimax* method

Course	Learning outcomes group	Component				
		1	2	3	4	5
Anatomy	A	0.518	0.098	-0.407	-0.276	-0.404
Physiology		0.542	-0.416	0.108	-0.332	-0.090
Microbiology		0.531	-0.128	0.220	-0.089	0.177
Parasitology		0.444	0.110	0.158	-0.121	0.512
Pharmacology		0.357	-0.598	0.260	0.149	-0.082
Psychology	B	0,486	0.301	0.012	0.057	-0.114
Pedagogics		0,566	-0.131	0.032	-0.199	0.064
ZdPublic health		0.280	-0.045	-0.577	-0.039	0.139
Foreign language		0.474	-0.359	0.150	0.094	0.159
Basics of maternity care	C	0.549	-0.151	0.150	-0.113	-0.125
Health promotion		0.393	-0.014	-0.391	0.191	0.565
Primary health care		0.455	-0.060	-0.257	0.471	-0.199
Obstetrical techniques and care during childbirth	D	0.618	-0.201	0.138	0.334	-0.124
Obstetrics and Maternity care		0.656	0.105	-0.185	-0.045	-0.259
Gynecology and Gynecological care		0.431	0.210	0.123	0.469	-0.212
Neonatology and Neonatal Care		0,511	0.306	-0.024	0.144	0.140
Paediatrics and Paediatric Nursing		0.448	0.077	0.130	-0.526	0.026
Internal medicine		0.492	0.372	-0.222	-0.027	0.168
Psychiatry		0.261	0.635	0.310	-0.171	-0.166
Rehabilitation in obstetrics, gynecology and neonatology		0.171	0.343	0.540	0.209	0.140

red indicates a clear solution
blue indicates an ambiguous solution

Source: author's own analysis

Criterion validity of the students' performance assessment system was checked by calculating Pearson's correlation coefficients between the grades for particular exam courses. The highest average correlation of grades was observed for *"Obstetrics and Maternity care"* ($r_{average} = 0.26$). As many as 13 courses did not reach average correlation coefficient value of over

0.20. The analysis of correlation gave particularly unfavourable results for "*Rehabilitation in obstetrics and gynaecology*" ($r_{average}$ = 0.07), "*Psychiatry*" ($r_{average}$ = 0.10), and *"Public health*" ($r_{average}$ = 0.10). The highest level of criterion validity was observed for the D-group courses ("Science in the field of specialist care"), with average values of Pearson's correlation coefficient (r) amounting to 0.20. See *Supplementary data* for a detailed breakdown of the results of the analysis of criterion validity.

Based on the course-related structure established in the standards of education, in which there are four groups of educational outcomes, evaluation of one-dimensionality of these groups was carried out to check whether each of them may be considered as one-dimensional. Principal component analysis was used to assess the eigenvalues and how a large part of variance was explained by the first component (**Table 5**). Only one eigenvalue for groups A and C was over 1, which, according to Kaiser criterion, showed one-dimensionality of each of them. But in groups B and D the first and the second variable exceeded the threshold and the percentage of the variance was below the assumed threshold of 40%.

Table 5. The share of variance explained by the first principal component

Learning outcomes group	Kaiser criterion	Participation of explained variance for the first factor (%)
A. Basic sciences	1.79	35.84
B. Social sciences	1.34; 1.09*	33.38**
C. Sciences in the basics of maternity care	1.36	45.45
D. Sciences in the field of specialist care	2.43; 1.10*	30.42**

***** *eigenvalue of factors 1 and 2 respectively*
****** *unfulfilled criteria of one-dimensionality*

Source: author's own analysis

Discussion

Measurement impartiality, i.e. independence of a measuring situation means providing all students with equal (fair) conditions for assessing their achievements. Creation of appropriate conditions during the exam and methods selection of students' achievements evaluation ensuring a comparable degree of independence concerning the measuring situation in consecutive years constitute an important element of measurement impartiality [11–13]. Impartiality of the assessment system is therefore the basis for a reliable and valid measurement of students' real achievements that allows for the evaluation of the quality of education at the faculty over subsequent years.

Reliability of measurement means reproducibility of results in specific conditions. Reliability is most frequently measured with the assessment of internal consistency of measurement results estimated on the basis of average variances of grades for all exams (*Cronbach's alpha)* [7]. The present results of reliability evaluation with Cronbach's alpha coefficient demonstrated that the students' performance assessment system had a sufficient level of reliability (Nunnally's criterion of $\alpha >$ 0.70 was met) [8]. Reliability decreased mostly due to a poor selection of examination methods and, in particular, the improper structure of test tasks that are the core of the assessment process. Students may not have an opportunity to show their achievements in a particular area if the exam significantly limits the learning content that is used to assess the outcomes of education. In addition, low reliability of the assessment system can be observed particularly in cases with a high degree of outcomes variations. This may happen both in general and specialist education where very diverse features and characteristics of students are evaluated.

The analysis of reliability for particular groups of educational outcomes demonstrated that the groups B and C represented a low level of internal consistency (alpha amounted to 0.438 and 0.397, respectively). This was largely related to a measurement scale, i.e. a small number of examination courses comprising groups B and C (four and three, respectively). This is a weakness of each measurement of reliability carried out with the use of Cronbach's alpha coefficient which is sensitive to the size of the measurement scale [9]. A large number of random errors in a particular measurement may constitute another reason for a low value of alpha coefficient. Random variations of assessment results accompany every measurement and directly affect the value of reliability coefficient [14]. At the value of $\alpha < 0.7$, random errors constitute more than 30% of the variability of results and, according to Guilford, measurement in such conditions might be used only for intergroup comparisons and not for individual differentiation [15]. In addition, it is worth noting that sole aspiration to obtain high values of alpha coefficient does not solve the problem of reliability since its large value only minimises the influence of random errors on the results. This, however, gives no certainty with respect to the presence of systematic errors (sometimes serious ones) relating to the measurement bias, which may be the case with each course or group of examination courses [15].

Apart from estimating the degree of internal consistency of the students' performance assessment system, it is important, for the evaluation of psychometric parameters of the measurement scale, to establish discrimination power for each of the examination courses. As defined by Brzeziński, discrimination power deter-

mines to what extent a result for one item in the scale (here: examination course) differentiates the students' population with respect to the feature it measures [16]. An analysis of discrimination index values for particular courses of four groups of educational outcomes led to a conclusion that the results for *Psychiatry* and *Rehabilitation in obstetrics and gynaecology* (indices of 0.193 and 0.143, respectively) fell far from general assessment of students' performance in this field of knowledge (specialist care in this case). Particularly high values of discrimination power were found for *Microbiology* (index of 0.439; group A), *Pedagogics* (index of 0.463; group B), *Basics of maternity care* (index of 0.444; group C), and *Obstetrics and Maternity care* (index of 0.562; group D). Exam results for these four courses constitute the core of the students' performance assessment system for particular groups of educational outcomes. It can be assumed that the principles of students' assessment for these courses may be treated as a model solution and should become the basis for the improvement of differentiation parameters for the remaining courses in a particular group of educational outcomes.

Apart from establishing the level of reliability of measurement, defining its validity is also important for the quality of assessment tools. Validity should be understood here as the degree of consistency to which a measurement tool measures what it is supposed to measure [16]. Therefore, we can speak of the effectiveness of a particular method for the assessment of a certain set of features and characteristics of a student [17]. There is no precise method used to measure validity; there is only its indirect assessment that may concern, among others, estimation of theoretical and criterion validity [4,6,16]. In the case of exams set out in the curriculum for Midwifery, content consistency of examination tasks with the objectives of education for particular courses is crucial for theoretical validity. In compliance with Kaiser criterion, the structure of thenstudents' performance assessment system at WUM took the 5-factor form, which is not in line with theoretical assumptions. In addition to that, the exploratory factor analysis showed that the distribution of students' grades for 20 courses within the structure of performance measurement scale explained only half of the general variability. Particular components were not homogeneous with respect to particular groups of courses. For instance, Component 1 comprised students' grades for educational outcomes from groups A, B, C, and D (13 courses in total). The analysis of one-dimensionality for each group of courses also demonstrated that in two cases the performance measurement scale was not homogeneous. This concerned courses from groups B and D. The results of estimating theoretical validity and one-dimensionality may lead to a conclusion that grades for particular courses did not reflect the division of educational outcomes into groups proposed in the standards of education for Midwifery.

Assessment of criterion consistency of students' results for particular examination courses was the other aspect of validity analysed in the present study. The correlation analysis demonstrated that, similarly to theoretical validity, assumptions for criterion validity were not met in all four groups of courses. The lowest values of correlation coefficients were found for the courses from group D, *Psychiatry* ($r_{average}$ = 0.10) and *Rehabilitation in obstetrics and gynaecology* ($r_{average}$ = 0.07) in particular. These results confirmed earlier findings on reliability and theoretical validity of the students' performance assessment system. With respect to criterion validity, students' grades for the courses from groups A and D ($r_{average}$ of 0.19 and 0.20, respectively) were the best, while grades for the courses in social sciences ($r_{average}$ of 0.16) produced the worst results.

The analysis of validity of educational measurement is supposed to prevent abuse in the interpretation of measurement results [18]. If a student gained a high grade point average, its value is important only when it reflects real student's achievements, particularly with respect to curricular requirements. Therefore, the estimation of prognostic validity relating to the assessment of to what degree the outcomes of education may serve to predict the future of students, e.g. their careers, constitutes an essential aspect of measurement quality analysis. The analysis of diagnostic features of educational measurement may give some more valuable indications as to the validity of assessment system. Thus, it is necessary to perform an external test to evaluate competence acquired by a student/graduate. If an obligation to pass a State Examination (an equivalent of the National Medical Exam) in order to obtain the right to exercise the profession is introduced for all graduates, it will be possible to verify the quality of the teaching process and diagnostic validity of the students' performance assessment system of a particular university. Analyses of the quality of teaching nurses in the US may serve as an example. American graduates need to pass the *National Council Licensure Examination for Registered Nurses* (NCLEX-RN), which constitutes an external verification of their preparation for the profession [19].

Conclusions

The reliability analysis of the achievements assessment system of Bachelor's degree students of Midwifery at Warsaw Medical University demonstrated that psychometric parameters in this respect were good. However, the analysis of validity raised serious concerns. In some cases, students' grades may not reflect their real fea-

tures and properties (vide *Psychiatry* or *Rehabilitation in obstetrics and gynaecology*). When quite good reliability goes hand in hand with insufficient validity, credibility of the entire competence assessment system decreases. This significantly influences the assessment of those students whose grade point average is near the bottom of the scale (grade point average < 3.0). Due to the fact that these are the weakest students, there is a serious risk of having a number of graduates with an insufficient level of output competencies.

References

1. 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).
2. Przewodnik dydaktyczny dla studentów kierunku położnictwo studia pierwszego stopnia. Warszawski Uniwersytet Medyczny, Warszawa 2013.
3. Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011; 33: 783–97.
4. Niemierko B. Diagnostyka edukacyjna. Warszawa: Wydawnictwo Naukowe PWN; 2009.
5. Norman GR, Vleuten C, Newble DI. International handbook of research in medical education. Dordrecht: Kluwer Academic Publishers; 2002.
6. Niemierko B. Pomiar wyników kształcenia. Warszawa: Wydawnictwo Szkolne i Pedagogiczne; 1999.
7. Feldt LS. A test of hypothesis that Cronbachs alpha or Kuder-Richardson coefficent 20 is same for 2 tests. Psychometrika. 1969; 34: 363.
8. Nunnally JC, Bernstein IH. Psychometric theory. 3 New York: McGraw-Hill; 1967.
9. Jankowski K, Zajenkowski M. Metody szacowania rzetelności pomiaru testem. W: Fronczyk K. (red.). Psychometria – podstawowe zagadnienia. Warszawa: Vizja Press & IT; 2009. 84–110.
10. Kaiser HF. The varimax criterion for analytic rotation in factor analysis. Psychometrika. 1958; 23: 187–200.
11. Rowley J. Measuring quality in higher education. Qual High Educ. 1996; 2: 237–55.
12. Tam M. Measuring Quality and Performance in Higher Education. Qual High Educ. 2001; 7: 47–54.
13. Sood R, Singh T. Assessment in medical education: Evolving perspectives and contemporary trends. Natl Med J India. 2012; 25: 357–364.
14. Niemierko B. Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe. 1st ed. Warszawa: Wydawnictwo Szkolne i Pedagogiczne; 1975.
15. Guilford JP. Psychometric methods. 2nd ed. New York: McGraw-Hill; 1954.
16. Brzeziński J. Metodologia badań psychologicznych. Warszawa: Wydawnictwo Naukowe PWN; 2002.
17. Goodwin LD. Changing conceptions of measurement validity: an update on the new standards. J Nurs Educ. 2002; 41: 100–106.
18. Kubielski W. Podstawy pomiaru, konstruowania i ewaluacji testu dydaktycznego. Warszawa: Wydawnictwo Wyższej Szkoły Pedagogicznej TWP; 2006.
19. Seldomridge LA, DiBartolo MC. Can success and failure be predicted for baccalaureate graduates on the computerized NCLEX-RN? J Prof Nurs. 2004; 20: 361–368.

The manuscript accepted for editing: 04.05.2016
The manuscript accepted for publication: 17.06.2016

Funding Sources: This study was not supported.
Conflict of interest: The authors have no conflict of interest to declare.

Address for correspondence:
Mariusz Panczyk
Żwirki i Wigury 61
02-091 Warsaw, Poland
phone: +48 22 57 20 490
e-mail: mariusz.panczyk@wum.edu.pl
Division of Teaching and Outcomes of Education
Medical University of Warsaw