Quality assurance in higher education: analysis of grades for reviewing course levels

Trudy Rexwinkel, Jacques Haenen & Albert Pilot

Quality & Quantity
International Journal of Methodology

ISSN 0033-5177
Volume 47
Number 1

Qual Quant (2013) 47:581-598
DOI 10.1007/s11135-011-9481-6

Springer

Qual Quant (2013) 47:581–598
DOI 10.1007/s11135-011-9481-6

RESEARCH NOTE

Quality assurance in higher education: analysis of grades for reviewing course levels

Trudy Rexwinkel · Jacques Haenen · Albert Pilot

Published online: 26 March 2011

Abstract In quality assurance, degree courses in European higher education have to demonstrate the course level by means of results. These courses produce many results, among them grades, referring to teachers' evaluations of students' performances, and it is our proposal to use them as an effective mean to reflect the course level. Our study examines the criteria that are needed to analyze grades as significant indicators for course levels. We considered what might constitute empirical proof of valid grades and from this analysis we established that the main proof of valid grades consists of measurements of construct validity, scale reliability, intercorrelation and face validity. The analyses delivered insights into the relationship between the proof of valid grades and elements of the curriculum. In the light of these insights we developed four characteristics as reference points for curricula realizing the course level. Following these points, we draw up a procedure to create such curricula. This procedure is explored in a study with eight bachelor degree courses. The conclusion is that the procedure traces the causes of invalid grades and confirms that valid grades are significant indicators of the course level.

Keywords Grades · Learning outcomes · Assessment · Validity · Reliability · Quality assurance

1 Do grades reflect the level of degree courses?

Currently, in international quality assurance in higher education, there is a growing focus on issues related to the complex nature and transparency of grades as a basis for the process of reviewing course level. The communiqués and decisions made by European ministers of

T. Rexwinkel (✉)
Faculty of Social and Behavioural Sciences, Utrecht University, Postbox 80127, 3508 TC, Utrecht, The Netherlands
e-mail: g.b.rexwinkel@uu.nl

J. Haenen · A. Pilot
IVLOS Institute of Education, Utrecht University, Postbox 80127, 3508 TC, Utrecht, The Netherlands

education since the Bologna Declaration stimulated an increased interest in European quality policy and the participating countries in the Bologna area are building this policy on common reference points. These developments have been strengthened with the establishment of the European Association of Quality Assurance (ENQA). Organisations and agencies of accreditation can become full members if the international panel has established the presence of the necessary infrastructure for independent and competent review. Auditors have to be able to review the institutions of higher education according to the internationally agreed Standards and Guidelines (ENQA 2009). These standards make it clear that empirical proof is required from institutions to demonstrate that they meet the standards. In this context the use of grades as an indicator for the course level could be useful. However, in degree courses it is often a problem to construct appropriate exams and award objective grades to students. For well-known reasons, the process of awarding grades is difficult: students experience difficulties in answering exam questions properly and accurately, and teachers often assess the first exam from a different perspective from that of the fifth, tenth or twentieth one. The assessor's attitude may be tolerant initially and become more rigorous, or vice versa, yielding various grades (Beran et al. 2009). These problems cause ambiguity in the proof of the grades.

Instead of the term grades, in the context of quality assurance the terms learning outcomes and learning results are used. Both notions are interpreted variously and are often confused with each other. Baume (2009, p. 26) conceptualises learning outcomes from the perspectives of aims as well as results. The classic learning outcomes describe what a student should be able to do in order successfully to complete a course of study. Learning outcomes are also conceptualised by results which have to be assessed. In the glossary of the Quality Assurance Agency for Higher Education (QAA 2010) the concept of outcomes is related to financial results. Following the description as being used in the Code of Practice for the Assurance of Academic Quality (QAA 2006), in this article we use the term grades to mean scores or numerical marks of students in higher education. This engenders the following problem: how can grades reflect the course level? This problem will be treated with the following research questions.

(1) How can the validity of grades empirically be proven?
(2) What criteria do curricula need to meet for reflecting the course level?
(3) What procedure and instruments can produce empirically proven grades?

To answer the research questions this article starts by examining the complexities of empirical proof for valid grades (Sect. 2). This proof lays the foundation for criteria (Sect. 3) and a procedure developing curricula that realize the course level (Sect. 4). This is explored in a case study with eight professional bachelor degree courses in physiotherapy (Sects. 5 and 6). Analysis of the data and the results (Sects. 7 and 8) are described and the article ends with conclusion and discussion (Sect. 9).

2 Analyzing grades

In this section we discuss how to analyze grades and to use them for insights into the strengths and weaknesses of the curriculum. Only analyzed grades can be an appropriate base for decisions concerning the curriculum. In answer to the first research question regarding empirical proof for valid grades it is assumed that analyzed grades are important indicators of the quality of the curriculum, particularly exams and evaluation. Valid grades indicate an appropriate curriculum. To examine this assumption grades are analyzed for validity and reliability with the following five techniques. First, the dataset is inspected on aspects of validity, since not

Table 1 Mean and SD from grades

Course subjects	M	SD
	Students ($n = 140$)	
Veterinary management	6.76	0.70
Anaesthesiology	6.99	0.88
General surgery	6.99	0.86
Veterinary public health	6.65	0.59
Veterinary medicine and society	7.45	0.69
General obstetrics	6.81	0.75
Clinical lectures	7.04	0.34
Average	6.96	0.69

every dataset is suitable for analysis. Second, construct analysis is selected as this technique answers the crucial question what is really measured (Gay and Airasian 2003, p. 139). This technique is relevant for the question how grades reflect the course level. Moreover, construct analysis informs us about patterns within the dataset. Third, the measured constructs are analyzed with the reliability of the scale (Field 2005, pp. 666–676). The structure in which reliability follows validity is in line with Moss's theory (1994). Fourth, Pearson correlation is measured, as the correlations demonstrate the relationships between the course subjects. Fifth, face validity refers to the acceptance of various stakeholders (Kane 2006, p. 36), since a study of the level of degree courses and grades includes the acceptance of various stakeholders. In order to illustrate our line of argument, the grades of two cases will now be presented and analyzed.

2.1 Valid grades

First, the dataset is inspected on aspects of validity. The data of the first case come from a scientific degree course in veterinary medicine. The grades belong to 140 students who graduated in 2004. In the Netherlands usually a 10-point scale is used ranging from one (very bad) to ten (excellent). The data are spread over seven course subjects; in four of these were no missing values (clinical lectures, obstetrics, veterinary management and anaesthesiology), in two course subjects (surgery and veterinary medicine and society) was 1 missing value and in one course subject 2 missing values (veterinary public health). The administration system of the grades follows the regulations students have to pass basic exams before they can do following, more complex exams and the rules referring to absence and illness of the students leading to a dataset including only the results of students who passed the exams in the same period. All these factors implicate a dataset suitable for analysis. The grades differ from 6 until 9 over seven course subjects as follows: for grade 6, meaning satisfactory, were in sum 274 scores, 28%; for grade 7, meaning more than satisfactory, were 492 scores, 50%; for grade 8, meaning good, were 190 scores, 20%; and for grade 9, meaning very good were 20 scores, 2%.

As can be seen in Table 1, *Mean* is on average 6.96 and $SD = 0.69$ indicating mean represents accurately the data. Noticeable is the small standard deviation of clinical lectures $SD = 0.34$ indicating that nearly all the scores for this course-subject are consistently close to the mean rating, implicating clinical reasoning runs along fixed schemes and protocols.

Second, the construct analysis of veterinary studies yields four constructs with ≥ 1 eigenvalue explaining 78% of the variance, which is a good percentage. The course subjects in veterinary management which have implications for the livelihood of farmers have the highest loadings on the first component. Veterinary medicine and society (referring to veterinary

Table 2 Construct analysis of grades

Course subjects	Components			
	Students ($n = 140$)			
	1	2	3	4
Veterinary management	**0.76**	0.22	0.08	0.07
Anaesthesiology	**0.76**	−0.07	0.26	0.28
General surgery	**0.70**	0.09	0.43	0.12
Veterinary public health	**0.62**	0.56	−0.15	−0.17
Veterinary medicine and society	0.08	**0.86**	0.27	0.20
General obstetrics	0.22	0.18	**0.88**	0.04
Clinical lectures	0.16	0.12	0.05	**0.95**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalisation
Bold indicates the highest course subjects' loadings on the respective components

legislation (national and EU) and animal welfare), general obstetrics and clinical lectures are separate components which have the highest loadings on the second, third and fourth components respectively (Table 2). Third, scale reliabilities of the first component, covering four items and 138 case numbers, give $\alpha = 0.75$. Reliability of all constructs, consisting of seven items and 138 case numbers, gives $\alpha = 0.77$, a good result (Committee on Test Affairs Netherlands (COTAN 2009).

Fourth, Pearson correlation analysis exposes the relationship between the course subjects. The correlations of veterinary medicine give a good picture of relationships between the course subjects. Most correlations are significant on two levels of probability. Clinical lecturers and general obstetrics correlate on the highest level of probability: $r = -0.21$, $p < 0.05$ and other course subjects correlate on the level $p < 0.01$. The course subjects general surgery and anaesthesiology indicate a close relationship $r = 0.57$, $p < 0.01$ reflecting real relations between surgery and anaesthesiology. Several subjects indicate a direct relationship with veterinary management such as general surgery $r = 0.48$, $p < 0.01$, anaesthesiology $r = 0.45$, $p < .001$, veterinary public health $r = 0.41$, $p < 0.01$ in terms of the business interest of the farmer. Only the course subjects veterinary public health and clinical lectures correlate $r = 0.09$ are not significant, indicating uncertainty about the meaning of the relationship. To understand the relationship between the course subjects and to explain the true situation, the correlations have to be significant (Table 3). The specialists on the degree course can discuss the correlations and decide whether to examine specific correlations in order to make them more or less strong.

Fifth, face validity refers to the agreement of various stakeholders. In this situation we use evaluation studies carried out by the external auditors in the context of quality assurance, more specific accreditation, the monitor of master students, the national survey for students, questionnaires from the institute of higher education unrolled at the degree course (see Sect. 5). In this case, the outcomes of these studies confirm the measurement results, signifying that the grades are proven to be sustainable and represent an appropriate curriculum.

Given all this, we may conclude that the outcomes of the statistic analysis of the data's inspection along the proposed line are reason for confidence in these grades. This first case study indicates that valid grades do indeed correspond with an appropriate curriculum. This analysis reveals: (1) the usability of the dataset and mean and *SD* indicate how well the mean

Table 3 Intercorrelations between grades

Course subjects	Students ($n = 140$)						
	1	2	3	4	5	6	7
1. Veterinary management		0.45**	0.48**	0.41**	0.30**	0.31**	0.23**
2. Anaesthesiology			0.57**	0.31**	0.24**	0.33**	0.29**
3. General surgery				0.34**	0.32**	0.40**	0.23**
4. Veterinary public health					0.30**	0.23**	0.09
5. Veterinary medicine and society						0.29**	0.22**
6. General obstetrics							0.21*
7. Clinical lectures							1

* Significant at $p < 0.05$; ** significant at $p < 0.01$

represents the data, (2) construct analysis answers the important question about what the grades really measure, (3) scale reliabilities of the constructs can confirm the power of this answer, (4) Pearson correlations are informative about the measured, objective, relationships between the course subjects, (5) in terms of face validity, evaluation studies explain the analysed outcomes and inform about the agreement of various stakeholders. More generally it can be concluded that the statistic analyses render objective proof of the validity of grades, corresponding with a curriculum without crucial problems.

2.2 Invalid grades

To interpret the quality of the grades of the course on veterinary medicine was not problematic. This ideal situation, however, is not usual. The second case treats a professional bachelor degree course in physiotherapy.

The set demonstrates 150 valid case numbers from 2009 from the final year of the course in physiotherapy. These grades indicate whether students graduate or not. The students receive the results (in text) in three categories: good, satisfactory or unsatisfactory. Underlying these categories are grades from the 10-point scale. The grades are spread over five course subjects as follows: there were 8 scores, 5%, for grades <6, meaning unsatisfactory; 121 scores, 80%, for grades 6 and 7, meaning (more than) satisfactory; 21 scores, 15%, for grades 8 and 9, meaning (very) good. The current grades are categorised in three scales, which results in a vague structure. Did the student fail with grade 5 or 2? Did the student pass with grade 6 or grade 7? There are no missing values, which means 100% of students participated in all course subjects. Nobody was ill. The administration system of the grades includes the final grades for exams which can be achieved in different years. Each year includes three exam periods including re-exam. These circumstances make it difficult to create a clear dataset. *SD* (1.40) is on average not a very accurate representation of the data (relatively high compared with the mean) (Table 4).

The construct analysis of physiotherapy delivers three constructs with ≥1 eigenvalue explaining 69% of the variance, which is a good percentage. The course subject theory has the highest loading on the first component; three practice-oriented subjects have the highest loadings on the second component and the subject presentations have the highest loading on the third component (Table 5). Note the low loadings generally, and specifically that of work placement. The scale reliability of the second construct ($\alpha = 0.22$) and of all constructs ($\alpha = 0.13$) are unsatisfactory results (COTAN 2009, p. 14). These reliabilities give grounds for concern.

Table 4 Mean and SD from grades

Course subjects	Students ($n = 150$)	
	M	SD
Theory	6.69	1.42
Work placement	6.91	1.42
Case-analysis	7.12	1.39
Practical exam	6.92	1.46
Presentations	6.99	1.33
Average	6.93	1.40

Table 5 Construct analysis of grades

Course subjects	Components		
	Students ($n = 150$)		
	1	2	3
Theory	**0.75**	0.32	0.02
Work placement	−0.74	**0.29**	−0.03
Case-analysis	0.23	**0.74**	−0.17
Practical exam	−0.26	**0.67**	0.21
Presentations	0.05	0.01	**0.97**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalisation
Bold indicates the highest course subjects' loadings on the respective components

Table 6 Intercorrelation between grades

Course subjects	Students ($n = 150$)				
	1	2	3	4	5
1. Theory		−0.20**	0.17*	0.01	0.02
2. Work placement			0.03	0.12	0.00
3. Case-analysis				0.11	−0.02
4. Practical exam					0.04
5. Presentations					

* Significant at $p < 0.05$; ** significant at $p < 0.01$

Pearson correlations of the relationship between theory and case analyses $r = 0.17$, $p < 0.05$ are at the highest level of probability (Table 6). The course subjects work placement and theory, $r = -0.20$, are significant at $p < 0.01$ even though negative. What subjects are learnt in the course subject theory? What are the theoretical issues in work placement causing negative relationships? The relationships between the other course subjects (ranging from $r = -0.02$ to $r = 0.12$) are not significant. It is not obvious what these correlations imply.

Evaluation studies, i.e. internal survey of the institute of higher education unrolled at the degree course, attest that students are not satisfied with the exam and assessment instruments. The exam assessments are ambiguous, as criteria are not used consistently and cause different assessments. Given all this, we conclude that the outcomes of analyses of scale reliabilities, correlations and face validity are not positive and give grounds for concern. This second case

proves invalid grades indicate problems in the curriculum. These analyses in the second case deliver similar insights to the first even though the results are less favourable.

2.3 Proof of the validity of grades

Both cases have been analysed with the same instruments and the same norms which distinguish the valid from the invalid grades. The first case indicates that valid grades indeed correspond with an appropriate curriculum. The second case proves that invalid grades correspond with problems in the curriculum. Both analyses illustrate that grades are important indicators of the quality of the curriculum. For these reasons grades have to be valid. The validity of grades is proven by the following criteria:

(1) Construct analysis measuring components with eigenvalues ≥ 1 explaining $\geq 50\%$ of the variance (Field 2005, p. 632; Maas 2009, p. 38).
(2) Scale reliability of the constructs measuring $\alpha \geq 0.60$ (Maas 2009, p. 14; Field 2005, pp. 666–676; COTAN 2009).
(3) Inter-correlation measuring relationships between the course subjects. The relationships between the course subjects should be significant on the levels $p < 0.05$ and $p < 0.01$ which are the highest levels of probability. High correlations of $p < 0.05$ indicate close relationships; low correlations represent no direct relationships and negative correlations indicate no relationships between course subjects (Field 2005, p. 140). Other relationships are not significant. The relationships between the course subjects need to be explained.
(4) Face validity measuring the feedback of various stakeholders such as students and teachers. The agreement is gleaned from evaluation studies and should be $>50\%$. For (pre) testing the data collection and processing (process efficiency) n ranges from 25 (small scale) to 100 (large scale) (Snijkers 2002).

These criteria answer the first research question. From these four measurements the validity of grades can be mapped out. What do these criteria mean for the arrangement of the curriculum? This is the second research question and we consider it now.

3 Characteristics curricula realizing the course level

Given the insights into the relation between valid grades and appropriate curriculum, we can act consistently. In answer to the second research question, we outlined features referring to the characteristics of appropriate curricula. The analyses of the cases demonstrated that invalid grades are linked to ambiguousness in exams and assessment instruments. These elements of the curriculum, however, are not isolated but cohere with other elements such as aims, educational opportunities, grades and regulations of exams. The following characteristics may be outlined as the basic elements of curriculum design.

(1) *Defined course level*. It is necessary to define the course level to establish the curriculum's base realizing the course level. For this aim a procedure is developed and validated. This had been outlined in detail elsewhere (Rexwinkel et al. 2011). In this article, we will shorter summarize the procedure defining the level of degree courses (DLDC), that results in basics that are tested on validity and reliability.
(2) *Variation of educational opportunities.* Educational opportunities have to be consistent with the basics of the degree course so that students learn the content as well as possible.

These opportunities or didactic forms are the settings created in the degree courses to enable students to learn the content with relevant learning activities.

(3) *Assessment of students* including compatibility of exams, objectifying assessment instruments and well-grounded scaling and grading. Exams have to be varied and compatible with the defined basics. *Compatible exams* suit the type of basics. *The exams have to be varied* to avoid possible system error. *Assessment instruments* have to be different and objectifying, including the process of awarding grades for student achievements. As it is complex for one individual teacher to assess stable, it is necessary to be aware of this and to use objectifying instruments as explicit criteria, answering models, blind assessment, more than one assessor as two assessors or a committee. *Well-grounded scaling and grading* implies that scales, norms, and meaning of the grades are established. Grades for students' achievements have to be awarded with well-grounded reasons. Clear explanations are based on elaborated assessment instruments and on the meaning of grades.

(4) *Regulations of exams* have to be established. The course has clear regulations covering student absence, illness and other circumstances. The aim of the regulations is to encourage all students to participate the exams in order to create a complete dataset that properly can be analyzed. These characteristics contribute to a curriculum producing grades that do not imply any crucial problems.

4 Procedure for developing curricula that realize the course level

Consequent on the characteristics a procedure is created to define the course level and develop curricula that realize this level. The aim is a feasible procedure that renders empirical proof. With this aim we address the third research question: what procedure and instruments can produce empirically proven grades?

Defining the course level refers to the procedure which results in basics that are tested on validity and reliability. The procedure is condensed reflected in Table 7. Core of it is to define the course level with content and that it is empirically proven. Systematically a frame of reference for content is created. Based on this content basics are formulated which have to meet various criteria referring to the elaborated complexities and characteristics of the course level. The resulting basics are unrolled in a survey at recent graduates. Following the steps of the DLDC-procedure validity is measured at various stakeholders and at independent experts. The resulting basics are tested on validity and reliability and lay a sound basis for the further development of the curriculum.

Developing the curriculum refers to the procedure resulting in an appropriate curriculum which is the condition for valid grades. Such a course of study consists of defined and validated basics, various educational opportunities, and assessment of students and regulations of exams that are consistent with the basics developed before. The curriculum is developed with the support of instrumentation and materials described in Sect. 5, and curriculum design and development (Stefani 2009). After the curriculum realizing the course level is completed, the procedure implies three rounds. In the first round, the curriculum's version one is developed and pretested on a small scale; in the second round version two is improved and tested on a larger scale and in the final round, version 3 is implemented at all students. Carrying out the procedure validity is measured at teachers, students and experts. The generated grades are tested on validity and reliability (Table 8).

Table 7 Condensed design of the procedure defining the level of degree courses (DLDC)[1]

Steps	Participant(s)	Instrumentation and materials
Stage 1. Creating a reference frame for content and aims. The course level is largely identified with content. This has to be selected and established as objectively as possible and also validated		
01 Elements of content and aims are outlined	Three to ten experts	With the support of questionnaire and diagram of association
Stage 2. Developing basics reflecting the course level. The validated content and aims are specified with basics indicating the core of the course level		
01 The content and aims are made specific with basics which have to meet criteria	Three to ten other experts	Criteria referring to complexities, characteristics, international understandable. Instruments for construct validity
Stage 3. Testing the defined course level. The basics are tested for validity and reliability		
01 The basics are screened and processed in a test instrument	Procedure manager	Awareness of ambiguity and constructing measuring instrument

[1]The complete DLDC-procedure takes 15 steps and includes measurements of validity and reliability

5 Method

5.1 Participants

In sum 394 persons participated in the exploratory case study: three experts on the content of the course, four independent experts in health care, graduates' coordinators of eight participating degree courses for professional bachelor degree courses in physiotherapy, coordinators of the eight degree courses, two members of the test office and 369 graduates participating in the survey.

5.2 Materials

Publications and evaluation studies are used as materials for our study. The publications concern particularly activities related to defining the course level. The materials were selected from different perspectives that had to cover the professional and educational perspectives of physiotherapy, as well as quality assurance. From the professional perspective we worked with profiles as *Professional Profile Physiotherapist*, published by the Royal Physiotherapy Organisation (KNGF 1998) with the text: and the Professional Profile Degree Courses Physiotherapy (Utrecht University 2004). From the perspective of education documents were adapted such as 'Competence profiles for degree courses in physiotherapy' (Utrecht University 2003). In the context of quality assurance we used reviews from accreditation organizations as the English 'Benchmark statement: Health care programmes' (QAA 2003) and the American Standards of Competence (FSBPT 2006).

Evaluation studies were used for the interpretation of the grades' analyses. These studies consist of accreditation reports of national accreditation organizations, national bachelor and master monitors carried out by research centres for education and labour market, surveys in the context of internal quality assurance executed by institutions of higher education and the degree courses.

Table 8 Procedure developing curriculum realizing the course level, extending the procedure defining the level of degree courses (DLDC)

Steps		Participant(s)	Instrumentation and materials
Stage 1. The main elements of the curriculum are developed in the first, improved in the second and implemented in the third round of the complete procedure			
01	Varied educational opportunities consistent with the basics are developed	Educational experts on educational opportunities	Constructs and basics as results of defining the course level
02	Various exams compatible with basics, different and objectifying assessment instruments and plan for awarding grades are outlined	Experts on constructing exams, assessment, course subjects, validity and exams	Accurate measurement of basics; instruments for stable assessment; scales, norms, meanings of grades, clear explanations
03	Regulations of exams covering student absence, illness and other circumstances are established	Experts on exams and administration	Planning of exams for the benefit of a proper dataset
Stage 2. Validating version 1 of the curriculum in the first round, version 2 and 3 in the second and third round			
01	The developed elements are described as a coherent curriculum	Procedure manager	Explanation and tables embedding the curriculum's elements
02	The curriculum is discussed	Teachers and students	Measurement face validity
03	The curriculum is validated	Experts on evaluating curriculum	Calculations of content validity
Stage 3. Pre-testing version 1 of the validated curriculum on a small scale (minimum 25) in the first round; testing version 2 on a larger scale (maximum 100) in the second; implementing version 3 at all students in the third round			
01	Version 1 is pretested with fixed group during fixed period	Teachers and students	Students pilot the validated curriculum in one semester
02	Evaluation studies on the curriculum are unrolled	Students, teachers in and outside the pilot	Questionnaires, interviews covering the complete curriculum
03	The data of evaluation studies are processed	Procedure manager	Analyses with measurements of validity and reliability
Stage 4. Analyzing and explaining grades of the versions as is carried out in the second stage			
01	Grades are analysed and explained	Procedure manager	Measurements of validity and reliability. Outcomes of other studies measuring the same entities with other instruments
02	Measures for improvement are developed and discussed	Teachers, students, experts on aspects of curriculum	Measurements of face validity
03	Measures are validated	Independent experts	Measurements of content validity

5.3 Procedure

The procedure is carried out in the exploratory case study explained in Sect. 6. The procedures defining the course level and developing the curriculum are detailed in the data analysis (Sect. 7) and results (Sect. 8).

6 Exploratory case study

To answer the final research question, this section describes the case study in which the procedure rendering empirically proven grades is explored through eight degree courses. The setting of the case study is eight professional bachelor degree courses in physiotherapy in the Netherlands in 2004. The new system of quality assurance was recently implemented and the coordinators of eight courses needed results reflecting the course level. This need implicated new issues including development of instruments, which were dealt with in an exploratory case study. This type of study is selected because problems with less familiar factors have to be solved. These studies are, according to Creswell (2007, p. 341), consistent with an exploratory case study design. The coordinators of the courses met regularly and comprised a rather strong group. The communication within this group was important for the success of the study and during the operation the course coordinators were consulted intensively and collaborated in the projected.

7 Data collection and analysis

The exploratory case study is an essential condition for this study because the profession of physiotherapy underwent developments which meant that less familiar issues had to be resolved.

7.1 Defining the course level

Creating a reference frame for content and aims required experts in physiotherapy and physiotherapeutic education to analyze themes, findings and developments in scientific disciplines and external surroundings which are meaningful for the degree courses. It was noticeable that most knowledge and policy came from the government's ministry for health care and the Royal Dutch Physiotherapy Organisation (KNGF). The themes were selected, discussed and agreed. Under the influence of changing governmental policy, the profession of physiotherapy changed. Physiotherapists were given additional new tasks besides purely physiotherapeutic treatments. This shift belonged to *holistic health management*, or health care focussing on the individual patient's needs. This development meant that professionals from several sectors and organizations had to retune their activities in order to create a coherent offer of care for the patient (integrated care). The profession of physiotherapist transformed *from repairer to coach* indicating the fewer use of passive techniques and a more active approach. *Prevention* also meant a reorientation from illness into health, involving more attention to stimulating healthy behaviour and responsibility for the citizen's own health. Activities such as advising, counselling and coaching became more important for physiotherapists. The physiotherapist could be directly consulted by patients; no longer was referral by a general practitioner necessary. This development of *free accessibility* meant greater autonomy of physiotherapists. It caused the shift to *evidence physiotherapy* including learning activities, such as

recognising the value of research and other scholarly activities in relation to the development of the profession and of patient / client care. These involved issues such as: gathering relevant information and adopting systematic approaches, using reasoning and problem-solving skills to make judgements in terms of prioritising actions and identifying risks.

Developing basics representing the course level: themes, findings and developments were specified into 53 basics meeting criteria related to the course level, referring to elements such as complexities of the course level, certain level characteristics, empirical proof and international comprehensibility (Rexwinkel et al. 2011; Koster et al. 2005).

Testing the defined course level. The agreed basics were unrolled in a survey of recent graduates of the eight professional degree courses. These basics were screened for ambiguousness and put into the measuring instrument. In compliance with the physiotherapeutic experts' advice the measuring instrument was constructed on a 4-point scale, ranging from 'not satisfactory' to 'more than satisfactory'. The numbers of the scale had normative meanings and were defined in terms with which students were familiar. The argument was that such a scale gives the user immediate normative information (Angoff 1971, p. 528). Data were collected and analyzed on validity and reliability. The eight coordinators of the degree courses approved the defined level.

7.2 Developing the curriculum that realizes the course level

The components of the construct analysis were the starting-point for the curriculum's development. The outcomes of the construct analysis are shown in Table 9. The coordinators of the degree courses agreed about creating a common model curriculum for all courses, that they elaborated in their separate individual courses in line with the policy of the various institutions. This common model is described in this section.

Educational opportunities such as problem-based learning (Chiou-Fen et al. 2010) were adopted by most degree courses in physiotherapy. The intention was students to learn structured solving problems derived from professional practice. The learning took place in the degree courses in various educational opportunities. Specific examples are skills-lab method, demonstrations with patients; simulation with patients, work placements, work experiences (Spreen et al. 2005); lectures by teachers, seminars, skills-training (Pool-Goudzwaard et al. 2005; Voskuil et al. 2005).

In the *skills-lab method* students exercised and trained their skills with actor-patients with feedback from teachers and fellow-students. The skills involved for example clinical reasoning, as physiotherapy is evidence-based and ground the therapies. Specific skills-training accommodate students to acquire less complex skills. Another type of educational opportunity was the *seminar* for groups of about twelve students. Cases as well as specific topics were discussed during the seminar. Seminars were used to train the capacity for reflection, the thinking through of specific problems. A well-known educational method was the *lecture method* that was appropriate for embedding and contextualising knowledge for larger groups of students. This method had to be used with care as it was only useful if it did not take too long and the presentation was attractive.

Exams included various types and alternative exams such as *portfolio* and *personal or professional development plan*. The basics were examined inside the degree course; exams in professional practice had to probe the learnt basics. *Case analysis* was appropriate for the development of and exam in reflection on the physiotherapeutic problems of a case and the disciplinary knowledge necessary for diagnosis. The *station exam* examined procedural skills of physical treatments and communication with patients. The *skills-lab exam* was intended for complex skills such as clinical reasoning. The student had to solve complicated cases.

Table 9 Model curriculum for professional bachelor degree courses physiotherapy

	Constructs and two basics	*N*	M	SD	*N* of basics	α	Variance[a] %	Educational opportunities	Exam	Assessment [b] instruments
13	Using theoretical knowledge of 1. Epidemiology 2. Methodology	363	2.62	0.85	2	0.92	2	Lectures Work placement	Skills-lab exam clinical reasoning Work placement exam (WPE)	Two assessors Explicit criteria Work placement committee (WPC)
11	Conceptual knowledge of various views on health 1. Differing visions on health 2. Several opinions on health care	363	2.97	0.69	3	0.85	2	Lectures Work placement	Knowledge exam	Two assessors Explicit criteria
01	Factual knowledge of law and rules referring to health care 1. Law for medical treatment's agreement 2. Registration council for professionals	362	2.35	0.82	6	0.89	28	Lectures Work placement	Knowledge exam	Committee prof. organization (CPO) Explicit criteria
02	Carrying out organisational aspects of the practice 1. Applying patient management 2. staff management	362	1.89	0.85	4	0.94	7	Lectures Skills-training	Station exam: administratively skills	Explicit criteria CPO
05	Influencing patient's behaviour 1. Altering patient's behaviour with learning processes 2. Modifying patient's behaviour with primary prevention	362	3.24	0.65	4	0.85	4	Lectures Skills-training	Station exam comm. skills WPE	Two assessors Explicit criteria WPC
07	Counselling patients and clients 1. Explicating the relation between patient's health and factors leading to health problems 2. Advising clients about measures to reduce health risk	363	3.67	0.50	3	0.86	3	Seminar Skills-training	Station exam comm skills WPE	Two assessors Explicit criteria WPC
03	Reflecting physiotherapeutic problems from various perspective 1. Thinking about physiotherapeutic themes from ethical standpoint	358	2.98	0.73	4	0.85	6	Seminar: discussing	Case-analysis	Two assessors Explicit criteria

Table 9 Continued

Constructs and two basics	*N*	M	SD	*N* of basics	α	Variance[a] %	Educational opportunities	Exam	Assessment[b] instruments
2. Discussing physiotherapeutic issues in international developments									
04 Diagnosing with knowledge of specific disciplines and subjects	363	2.98	0.72	7	0.79	5	Seminar: thinking through	Case-analysis	Two assessors
1. Diagnosing with knowledge of pathology									Explicit criteria
2. Analysing patient's problem with knowledge of neurology									WPC
06 Integrating developments into professional practice	353	3.05	0.77	4	0.80	4	Skills-lab method	Skills-lab exam	Two assessors
1. Mixing patient's perspective (individualising) into professional practice								clinical reasoning WPE	Explicit criteria WPC
2. Adopting new treatment methods in professional practice									
08 Diagnosing physiotherapeutic needs	346	3.03	0.72	5	0.74	3	Skills-lab method	Skills-lab exam	Two assessors
1. Determining if physiotherapeutic treatment is necessary								clinical reasoning	Explicit criteria
2. Identifying the problem in line with the professional council's guidelines								WPE	WPC
10 Evaluating patient's progress on the basis of evidence	362	3.21	0.67	5	0.80	2	Skills-lab method	Skills-lab exam	Two assessors
1. Evaluating conscientiously with scientific evidence								clinical reasoning	Explicit criteria
2. Analysing systematically patient's treatment								WPE	
12 Judging professional literature	363	3.03	0.80	3	0.76	2	Seminar: thinking through	Methodological exam	Two assessors
1. Judging professional literature in English language									Explicit criteria
2. Evaluating professional literature on methodological quality									Blind assessment
09 Leading physiotherapeutic practice	352	2.34	0.94	3	0.92	3	Self-study	WPE	Explicit criteria
1. Complying with the law on physiotherapy									CPO
2. Achieving financial targets									

[a] The data of the first eight columns are explained in Sect. 8.1 [b]The outcomes of the final three columns are detailed in Sect. 7.2

These skills were also measured in *work placement exams (WPE)*. The basics of managing a physiotherapeutic practice were examined by a *committee of the professional organisation (CPO)* that tended to examine all basics with knowledge tests; however it was more efficient to measure some basics in work placement exams.

Assessment instruments have to be objectifying as one has to be aware of the difficulty of an individual teacher in assessing consistently and objectively. For the use of peer, self- and co-assessment Dochy et al. (1999) gave supporting oversight. Various exams were best assessed by *two assessors* supported by assessment instruments such as *explicit criteria* or *answer models*, which also had to be developed and (pre)tested before complete implementation. The two assessors were not fixed pairs, but varied. The explicit criteria were also used by a *work placement committee* who audited the assessment of the external (bachelor degree) teachers. The basics of leading a physiotherapeutic practice were assessed by the *committee of professional organisations* using explicit criteria.

Scaling and grading implied that scales, norms and meaning of grades were agreed by the coordinators of the degree courses. They agreed with the four-point scale with the normative meanings known in their degree courses. They agreed with the general norms for measurements in the context of quality assurance: recent graduates had to master the basics more than satisfactorily, and students in the final study year (one study year takes 60 credits measured in European credit transfer system, ECTS; these credits reflect the duration of the study) had to master the same basics satisfactorily; students had to master during the level-in-development (60 or 120 ECTS) the final basics nearly satisfactorily and for students in the first study year (60 ECTS) the basics were measured to map out the initial situation. The grades should award well-grounded to student's achievements with the knowledge of the (validated) basics, exams, assessment instruments, scales, norms and meaning of grades.

8 Results

8.1 Defining the course level

The frame of reference was created and resulted in nine main themes, findings and developments. These were discussed by the group of experts and the coordinators of the courses. The outcomes were measured with face validity, referring to the acceptance of various stakeholders. The positive agreements were counted. The three experts and eight coordinators of the courses agreed on seven themes, or 75%, which represented good agreements.

Fifty-three basics specifying the themes were discussed in the expert group and in the group of coordinators of the eight degree courses. The content validity of the basics was measured with Cohen's Kappa. Independent experts evaluated 53 basics, resulting in 47 agreements and Cohen's Kappa = 0.79, which were good agreements.

The defined level was tested in a survey of the basics at recent graduates of the eight bachelor degree courses. *Representativeness* was indicated by the sample size and the respondents' characteristics. In this test, 364 graduates of eight courses participated. This number agrees with the number of a quantitative pilot study calculated by Snijkers (2002, p. 65). The respondents had a fairly good spread over the eight degree courses (about 45 respondents per degree course). Of these graduates, 119 graduated in 2004 (33%); 126 of them in 2003 (35%); 74 in 2002 (20%); 41 in 2001 (11%) and three (0.8%) in 2000. Although the target group actually consisted of graduates of the last three cohorts, we kept them all in the sample, because the exploratory case study also aimed to develop instruments. Of the respondents 112 were males (31%) and 251 were females (69%); 170 (63%) worked in a physiotherapeutic practice, 23

(9%) in a hospital, 22 (8%) in a nursing home, 21 (8%) in a rehabilitation centre, one (0.4%) in a preventative centre and 34 (13%) elsewhere. The sample also matches the group's characteristics and is representative as regards development of instruments. In the context of a pilot study the numbers of graduates in the subsamples are representative; mean on average = 2.87 and rather high, $SD = 0.75$ which is rather large, indicating SD does not represent the data very accurately.

Construct analysis is the most important analysis, as it focuses on the question of what is really measured (Harkness et al. 2003). The outcomes of the construct analysis of the basics consisted of thirteen constructs with eigenvalue ≥ 1, explaining 71% of the variance all with good scale reliabilities. These results are shown in Table 9.

8.2 Developing the curriculum that realizes the course level

The specific educational opportunities and exams, which had to be compatible with the basics, and the assessment instruments, that had to be different and objectifying, are shown in the model curriculum for professional bachelor degree courses physiotherapy (Table 9).

As it became clear that well-considered (strict) regulations referring to absence and illness during exams were missing in most cases, the coordinators agreed to revise the current regulations.

The described curriculum was discussed by the coordinators of the eight degree courses in the second stage. Six coordinators agreed with the complete model (75%). The content validity was measured with interrater reliability by two independent experts, one in education and one in physiotherapy. The measurement of the model curriculum containing 13 constructs, 53 basics cohering with various and compatible educational opportunities, exams, assessment instruments and regulations of exam yielded Cohen's Kappa = 0.89, which indicated very good agreements (COTAN 2009).

9 Conclusions and discussion

In this article the validity of grades is examined in terms of three research questions. The first question concentrates on identifying what is needed for the empirical proof of valid grades. To answer this question, the grades of two degree courses were analysed in five aspects: first, the usability of the data; second, construct analysis to establish what the grades were really measuring; third, the reliability of the measured components; fourth, the correlations, as these render objective information about the relationships between the course subjects; fifth, face validity with measurements from various stakeholders. These analyses delivered important insights in the characteristics of curricula. The second research question concerns criteria for curricula that realize the course level. Based on the developed comprehensions we developed four characteristics focusing on defining the course level, variation of educational opportunities, assessment of students and the exam regulations. The third research question refers to the procedure and instruments which produce empirically proven grades. This question was answered by the procedures of defining the course level and developing curricula that realize the course level, as carried out in an exploratory case study of eight bachelor degree courses. It resulted in an empirically proven and defined course level and an agreed model curriculum for the eight degree courses. We conclude that the procedure traces the causes of invalid grades and confirms that valid grades include sustainable proof for grades as results reflecting the course level.

While we were developing this procedure and researching it, some points of reflection emerged. To some extent these are related to the European dimension in which empirical proof is the internationally understood language. The degree courses produce this evidence and the auditors of accreditation organisations or agencies review it. The ENQA reviews on outcomes report that in some countries it is a real problem to meet the standards. We refer to three cases. In case A it appears that expertise about evaluation is not present (Sandahl et al. 2006, p. 9). In case B subjective assessments of programme content, staff quality and traditional input measures form the basis of the judgements. The ENQA-panel, consisting of international members, meeting the criteria of ENQA and reviewing national organizations and agencies for accreditation on request of ENQA indicates it is a challenge to systematically assess student learning and use such empirical evidence to guide their efforts to improve quality. Without this type of concrete evidence at the institutional level alternative efforts to improve academic quality are likely to be wasteful and ineffective (Konrad et al. 2007, p. 23). In case C a substantial part of the quality consists of 'soft' instruments. The panel would welcome a more systematic approach. Moreover, international experts are not represented in the panels in case C. The number of themes, criteria and points of attention in the procedures should be reduced and the agency should be transparent about the criteria used. The duration of the entire process should be shortened and the internal accountability procedures should be improved (Crochet et al. 2009, p. 6, 20, 35).

In this article we presented a procedure producing empirical proof for grades that are indicators of the course level. Working with the proposed procedure offers valuable opportunities for comparing the level of courses in an international context. We consider our work to be a contribution to the ongoing debate on the usability of grades as an internationally understood quality assurance that a course is fulfilling its basic demands. This is of the utmost importance to produce assessors who are more confident about their assessment, students who understand better the grades they receive, and educational management who can better and more efficiently prepare the process of quality assurance.

Acknowledgements We thank the participating coordinators, teachers and students who remain anonymous but without whose contribution there would be no article.

References

Angoff, W.H.: Scales, norms and equivalent scores. In: Thorndike, R.L. (ed.) Educational Measurement, pp. 508–600. American Council on Education, Washington (1971)

Baume, D.: Writing and using good learning outcomes. Leeds Metropolitan University. http://www.leedsmetac.uk (2009). Accessed 15 Sep 2009

Beran, T., Violato, C., Kline, D., Frideres, F.: What do students consider useful about student ratings?. Assess. Eval. High. Educ. **34**(5), 519–527 (2009)

Chiou-Fen, L., Meei-Shiow, L., Chun-Chih, Ch., Che-Ming, Y.: A comparison of problem-based learning and conventional teaching in nursing ethics education. Nurs. Ethics **17**(3), 373–382 (2010)

Commissie Testaangelegenheden Nederland van het Nederlands Instituut van Psychologen/NIP [Committee on Test Affairs Netherlands] (COTAN).: Beoordelingssysteem voor de kwaliteit van tests. [System of assessing the quality of tests] http://www.psynip.nl (2009). Accessed 10 Oct 2009

Creswell, J.W.: Qualitative Inquiry and Research Design. Choosing Among Five Approaches. University of Nebraska-Lincoln, Lincoln (2007)

Crochet, M., Harvey, L., Toit Du, H., Sursock, A., Douterlungne, M., De Fraeye, G., Hover, C.: Report of the committee of the review of The VLIR Quality Assurance Unit. European Association of Quality Assurance (ENQA). http://www.enqa.eu/pubs_review.lasso (2009). Accessed 29 Mar 2010

Dochy, F., Segers, M., Sluijsmans, D.: The use of self-, peer- and co-assessment in higher education. Stud. High. Educ. **24**(3), 331–350 (1999)

European Association of Quality Assurance (ENQA). Standards and Guidelines for Quality Assurance in the European Higher Education Area, 3rd edn. http://www.enqa.eu/pubs_lasso (2009). Accessed 29 Oct 2009

Federation of State Boards of Physiotherapy. Standards of competence. http://www.FSBPT.org/download/StandardsOfCompetence (2006). Accessed 19 Sep 2010

Field, A.: Discovering Statistics using SPSS. Sage, London (2005)

Gay, L.R., Airasian, P.: Educational Research. Competencies for Analysis and Applications. Merill Prentice Hall, (2003)

Harkness, J.A., Van de Vijver, F.J.R., Mohler, P.Ph.: Cross-cultural survey methods. Wiley, Hoboken (2003)

Kane, M.Y.: Validation. Educ. Meas. **4**, 17–64 (2006)

Koster, B., Brekelmans, M., Korthagen, F., Wubbels, Th.: Quality requirements for teacher educators. Teach. Teach. Educ. **21**, 157–176 (2005)

Koninklijk Nederlands Genootschap voor Fysiotherapie (KNGF).: Beroepsprofiel Fysiotherapeut. KNGF. Bohn Stafleu Van Loghum, Amersfoort. [Royal Dutch Physiotherapy Organization: Professional Profile Physiotherapist.] (1998)

Konrad, H., Leynse, F., Crochet, M., Sursock, A., Campbell, C., Dill, D., Neetens, S., Hover, C.: Report of the committee for the review of the Accreditation Organization of The Netherlands and Flanders (NVAO). European Association of Quality Assurance (ENQA). http://www.enqa.eu/pubs_review.lasso (2007). Accessed 29 Oct 2009

Maas, C.: Validity: Factor-analysis. Grondslagen Psychologische diagnostiek en testtheorie. Cursusjaar 2009–2010. [Foundations Psychologic Diagnostics and Test Theory. Course year 2009–2010], pp. 7–53. Utrecht University, Faculty of Social Sciences, Utrecht (2009)

Moss, P.A.: Can there be validity without reliability? Educ. Res. **23**(2), 5–12 (1994)

Pool-Goudzwaard, A.N., Eaglestone, A., Van Raaijen, W.G., Broers, G.: Adviesrapport Accreditatie HBO bachelor opleiding Fysiotherapie [Advisory report accreditation professional bachelor degree course], voltijd, Hogeschool van Utrecht, Faculteit Gezondheidszorg. Hobéon Certificering BV. http://www.nvao.net (2005). Accessed 10 Apr 2009

Quality Assurance Agency for Higher Education (QAA).: Benchmark statement: health care programmes. http://www.quaa.ac.uk (2003). Accessed 15 Feb 2004

Quality Assurance Agency for Higher Education (QAA).: Code of Practice for the Assurance of Academic Quality and Standards in Higher Education, 2nd edn. Section 6 Assessment of Students. http://www.quaa.ac.uk (2006). Accessed 17 Jul 2007

Quality Assurance Agency for Higher Education (QAA).: Acronyms and glossary of main terms. http://www.qaa.ac.uk/aboutus/acronyms.asp (2010). Accessed 20 Jul 2010

Rexwinkel G.B., Haenen J.P.P., Pilot A.: Defining the level of degree courses. Manuscript submitted for publication (2011)

Sandahl, R., Rohlin, M.,Waerness, M., Brennan J., Sjunesson, K.: (2006). Evaluation of the Swedish National Agency for Higher Education. European Association of Quality Assurance (ENQA). http://www.enqa.eu/pubs_review.lasso (2006). Accessed 29 Oct 2009

Snijkers, G.J.M.E.: Cognitive laboratory experiences: on pretesting computerised questionnaires and data quality. Doctoral Dissertation, Utrecht University. Centraal Bureau voor de Statistiek [Central Office Statistic Netherlands] (2002)

Spreen, A.E, Voskuil, E.J., Boekholt, N.S., Van Beers, M.J.J.: Hogeschool Zuyd, Heerlen, Fysiotherapie, hbo-bachelor [report accreditation professional bachelor degree course physiotherapy], voltijd. Netherlands Quality Agency. http://www.nvao.net (2005). Accessed 10 Apr 2009

Stefani, L: Planning teaching and learning: curriculum design and development. In: Fry, H., Ketteridge, S., Marshall, S. (eds.) A Handbook for Teaching and Learning in Higher Education. Enhancing Academic Practice, pp. 40–57. Routledge, New York (2009)

Utrecht University.: Competentieprofielen Opleidingen Fysiotherapie. IVLOS, Institute of Education, Utrecht [Profile of Competences Degree Courses Physiotherapy.] (2003)

Utrecht University.: Beroepsprofiel Fysiotherapeut. IVLOS, Institute of Education [Professional Profile Degree Courses Physiotherapy] (2004)

Voskuil, E.J., Boekholt, N.S., Veen Van der, L.S., Van Beers, M.J.J.: Hanzehogeschool Groningen, Fysiotherapie, bachelor [report accreditation professional bachelor degree course physiotherapy], voltijd, Netherlands Quality Agency, http://www.nvao.net (2005). Accessed 10 Apr 2009