Pharmacy Education, 2010; 10 (2): 119-128

Difficulty and discrimination indices as quality assurance tools for assessments in a South African problem-based pharmacy programme

SOPHIA FOURIE[1], BEVERLEY SUMMERS[2], MONIKA ZWEYGARTH[2]

[1]*Department of Pharmacology, University of Limpopo, Medunsa Campus, Box 225, Medunsa, 0204. South Africa.*

[2]*Department of Pharmacy, University of Limpopo, Medunsa Campus, Medunsa, 0204. South Africa.*

Abstract

This study investigated the difficulty and discrimination ability of examination questions in an undergraduate pharmacy programme presented at the Medunsa Campus of the University of Limpopo, South Africa. This investigation was part of an education study, which evaluated the quality of knowledge assessments in this outcomes and problem-based BPharm programme. Indices of difficulty (where a higher index characterises an easier item) and indices of discrimination were calculated for each True/False item and constructed response question from a total of 15 summative examinations in the first- to fourth-year level. We adapted the item analysis calculation methods to additionally accommodate questions counting more than one mark. Mean difficulty indices were 66.7% for True/False items without negative marking; 55.1% with negative marking, and 60.1% for constructed response questions. Discrimination indices for True/False items were 0.22 without negative marking, 0.24 with negative marking, and 0.28 for constructed response questions. Factors are discussed which potentially influence the difficulty and discrimination of examination items and the usefulness of item analysis techniques to review and improve future assessments in this programme.

Keywords: *Item analysis, difficulty, discrimination, assessment, test construction, evaluation, item bank*

Introduction

Outcomes-based and problem-based learning (PBL) approaches focus on developing the competence of students. Acquisition and assessment of scientific knowledge, underpinning all basic and applied pharmaceutical sciences, is essential to achieve pharmacy competency outcomes. Assessment of student performance and evaluation of the assessment process itself, play an important role in the establishment of the quality of academic programmes (Anderson, 2005).

The BPharm programme, collaboratively presented by the Schools of Pharmacy of the Medunsa Campus, University of Limpopo (UL) and Tshwane University of Technology (TUT) since 1999, is the first South African pharmacy programme to follow an integrated, modular, themes-based, outcomes-based (OBE), problem-based (PBL), and experiential learning approach (Summers *et al.,* 2001). Outcomes from knowledge, skill and attitude domains of learning are assessed by a wide variety of continuous and summative assessments. This study focuses on the assessment of knowledge as conducted in the "traditional" written papers, which students write at the end of each study module and semester of the BPharm degree.

As part of the institution's quality assurance programme, the quality of these BPharm assessments was investigated. The basis of quality assurance in assessments normally includes characteristics such as the validity and reliability of questions and examinations (Linn & Gronlund, 2000). In addition we also investigated the appropriate difficulty, discrimination, and depth of knowledge (Bloom's Taxonomic level) of the questions (Fourie, 2004). This study of Fourie used a number of quantitative methods to investigate knowledge assessments in the programme and evaluated them in terms of each of the five characteristics of good assessment mentioned. Content validity was measured by matching the questions with learning objectives and face validity by means of a student opinionnaire. Reliability was measured by two methods: internal consistency was determined by Cronbach's Alpha and reproducibility was evaluated with the Equivalent forms approach. The depth of knowledge of questions was analysed across the six levels of Bloom's Taxonomy (Bloom *et al.,* 1956; Fourie, 2004)

*Correspondence: Sophia Fourie, *Department of Pharmacology, Medunsa Campus, University of Limpopo, Box 225, Medunsa, 0204. South Africa. Tel: +27 125214974. Fax: +27 125214121. E-mail: sfourie@ul.ac.za*

ISSN 1447-2701 online

This manuscript reports on a part of the main study, the determination and analysis of the difficulty and the power of discrimination of questions. The aim of this manuscript was to discuss use of item analysis methods for test construction and item banking and the contribution of these methods to evaluate the educational quality of questions and examinations in a pharmacy programme.

Item analyses techniques: Difficulty and discrimination indices

The difficulty index of an item (question) estimates the proportion of students who answered the question correctly (McCown *et al.,* 1996). It is usually expressed as a percentage ranging from 0% to 100%. A higher percentage characterises an easier test, which means that one can actually think of the difficulty index as an '*easiness index'*.

The discrimination index of an item measures its effectiveness to separate the high achievers in a group from the low achievers (Aiken, 1982), or distinguish between examinees who are knowledgeable or not (Professional Testing, 2008). It ranges from –1 to +1. An item discriminates positively if more students in the high-scoring group than in the low-scoring group answer it correctly; and opposite results give a negative discrimination. Both the difficulty and discrimination indices of a test are dependent on the composition and characteristics of the group to which the test is administered.

The calculation methods for the indices of difficulty and discrimination include only the scores of the highest and lowest groups of performers in a test. This is a method widely used, but educators use dissimilar percentages to divide the classes in groups, it varies from 25% to 50%. Linn & Gronlund (2000), and McCown *et al.* (1996) recommend 25%; Ebel & Frisbie (1991) and Dreckmeyer & Fraser (1991) suggest 27%; Singh *et al.* (2003) used 30% or 50%, depending on the size of the class. Ebel and Frisbie (1991) preferred 27%, but found no statistical difference between 25%, 27% or 33%.

Methods

This study was a retrospective, non-experimental quantitative evaluation. Student marks from True/False questions were manually captured from individual student answer papers, and marks from constructed response questions were retrieved from the departmental database. The questions were analysed by the item analysis techniques described below.

Study site

Although the BPharm programme is collaboratively presented by two Schools of Pharmacy situated on two campuses, this study was limited to examinations at the Medunsa Campus.

Assessments and questions

Fifteen final summative assessments were included in the study, seven End of Module (EOM) and eight End of Semester (EOS) examinations. All examinations consisted of two sections. Section B had 8 to 20 constructed response questions of varying length, including calculations and cases etc. Section A had a selected-response item format with 80 to 160 True/False items, grouped into sets of four (see Appendix II for examples). In the examinations negative marking was applied with the True/False (T/F) questions to discourage students from guessing. For the purpose of the study, student response to each T/F item was manually captured as a separate item and not as a set. No negative marks were subtracted in the analysis.

Student classes and students

This study included all the examinations administered from the first to fourth year level in 2001 and 2002, as well as the 2003 examinations of the fourth years - to provide the fourth years with two sets of examinations.

The results of all students from the above mentioned examinations were used, excluding the results of nine students who failed any year, to avoid evaluation of the same student twice on the same work.

Analysis method

This study used the method of Ebel and Frisbie (1991) as applied by McCown and co-workers (1996) to calculate indices of difficulty and discrimination of the T/F items.

- For each examination paper, each student's marks for each single T/F item were manually captured on Microsoft Excel from the hard copies of the student answer papers.
- The mean marks for the students for the T/F items in the A section were sorted from the highest to the lowest.
- The student marks were divided in 4 quarters and the marks of the two middle groups deleted.
- Only marks achieved by the 25% highest and 25% lowest performers were used for the two equations below.
- The top and bottom sections must have an equal number of students.

$$\text{Difficulty index} = p$$

$$\frac{\text{Number of correct responses by high-scoring group} + \text{Number of correct responses by low-scoring group}}{\text{Total number of students in the two groups}}$$

$$\text{Discrimination index} = d$$

$$\frac{\text{Number of correct responses by high-scoring group} - \text{Number of correct responses by low-scoring group}}{\text{Number of students in one of the two groups}}$$

The two item analysis equations above are written for one mark/binary questions only. Therefore the equations were modified for questions in Section B, which count more than one mark (Fourie, 2004).

Difficulty [or Discrimination] index (p or [d])

$$\frac{\text{Marks scored by high-scoring group} + [-] \text{Marks scored by low-scoring group}}{\text{Total number of students in the two groups [in one of the groups]} * \textbf{maximum possible marks}}$$

In addition to indices for single items or questions, the following mean difficulty and discrimination indices were calculated and compared: per examination, per year group, for each section: i.e. Section A without negative marking, Section A with negative marking, and Section B. Means for the different sections (n=15 examinations) were compared.

Statistical analysis

Student's t-tests and ANOVA were performed on SAS or MS-Excel to determine whether the differences obtained in the test results were significant. The level of significance was $p \leq 0.05$.

Ethical approval: Ethical approval was granted by the Research, Ethics and Publications Committee of the University's Faculty of Medicine. Consent was obtained from the BPharm students to use their examination results for research.

Results

A total of 15 examinations in the pharmacy programme was analysed, three to four examinations from each academic year group (see Tables 1 and 2).

Table 1:
Difficulty indices (p) of T/F items in EOS and EOM examinations

Students and Examinations			Difficulty index as %	
Year	Number of T/F items per paper	Number of students	Mean p per paper	(SD)
End of Semester (EOS) Examinations				
1st year students				
2001	140	40	65.3%	(2.0%)
2002	160	54	68.7%	(1.8%)
Mean values: 1st years			**67.0%**	
2nd year students				
2001	160	30	63.3%	(2.0%)
2002	160	40	65.6%	(1.9%)
Mean values: 2nd years			**64.5%**	
3rd year students				
2001	160	28	67.5%	(2.0%)
2002	160	30	71.6%	(1.9%)
Mean values: 3rd years			**70.0%**	
4th year students				
2002	160	28	65.0%	(2.1%)
2003	144	29	61.6%	(2.3%)
Mean values: 4th years			**63.3%**	
Overall mean in 8 EOS examinations			**66.1%**	
End of Module (EOM) Examinations				
1st year students, Module 1.6				
2001	140	40	62.5%	(2.0%)
2002	120	54	68.3%	(2.0%)
Mean value: 1st years			**65.4%**	
2nd year students, Module 2.3				
2001	100	30	60.4%	(2.3%)
2002	100	40	65.7%	(2.3%)
Mean value: 2nd years			**63.1%**	
3rd year students, Module 3.5				
2001	144	28	73.5%	(2.0%)
Mean value: 3rd years			**73.5%**	
4th year students, Module 4.2				
2002	120	28	71.9%	(2.3%)
2003	100	30	70.4%	(2.5%)
Mean value: 4th years			**71.2%**	
Overall mean: 7 EOM papers			**67.5%**	
Overall mean: 15 examinations			**66.7% ± 3.9%**	

Table 2:
Discrimination Indices (d) of T/F items in EOS and EOM examinations

Students and Examinations			Discrimination Index	
Year	Number of T/F items per paper	Number of students	Mean d per paper	(SD) per paper
1st year students				
2001	140	40	0.24	(0.02)
2002	160	54	0.22	(0.02)
Mean value: 1st years			**0.23**	
2nd year students				
2001	160	30	0.25	(0.02)
2002	160	40	0.25	(0.02)
Mean value: 2nd years			**0.25**	
3rd year students				
2001	160	28	0.22	(0.02)
2002	160	30	0.18	(0.02)
Mean value: 3rd years			**0.20**	
4th year students				
2002	160	28	0.16	(0.02)
2003	144	29	0.26	(0.02)
Mean value: 4th years			**0.21**	
Mean of 8 EOS examinations			**0.22**	
1st year students, Module 1.6				
2001	140	40	0.23	(0.02)
2002	120	54	0.24	(0.02)
Mean value 1st years:			**0.23**	
2nd year students, Module 2.3				
2001	100	30	0.28	(0.02)
2002	100	40	0.28	(0.02)
Mean value 2nd years:			**0.28**	
3rd Year EOM 3.5				
2001	144	28	0.21	(0.02)
Mean value 3rd years:			**0.21**	
4th year students EOM 4.2				
2002	120	28	0.20	(0.02)
2003	100	30	0.21	(0.02)
Mean value 4th years			**0.20**	
Mean of 7 EOM examinations			**0.24**	
Overall mean of 15 examinations			**0.23**	**Mean individual items 0.22 (0.04)**

Table 1:
Difficulty indices (p) of T/F items in EOS and EOM examinations

Students and Examinations			Difficulty index as %	
Year	Number of T/F items per paper	Number of students	Mean p per paper	(SD)
End of Semester (EOS) Examinations				
1st year students				
2001	140	40	65.3%	(2.0%)
2002	160	54	68.7%	(1.8%)
Mean values: 1st years			**67.0%**	
2nd year students				
2001	160	30	63.3%	(2.0%)
2002	160	40	65.6%	(1.9%)
Mean values: 2nd years			**64.5%**	
3rd year students				
2001	160	28	67.5%	(2.0%)
2002	160	30	71.6%	(1.9%)
Mean values: 3rd years			**70.0%**	
4th year students				
2002	160	28	65.0%	(2.1%)
2003	144	29	61.6%	(2.3%)
Mean values: 4th years			**63.3%**	
Overall mean in 8 EOS examinations			**66.1%**	
End of Module (EOM) Examinations				
1st year students, Module 1.6				
2001	140	40	62.5%	(2.0%)
2002	120	54	68.3%	(2.0%)
Mean value: 1st years			**65.4%**	
2nd year students, Module 2.3				
2001	100	30	60.4%	(2.3%)
2002	100	40	65.7%	(2.3%)
Mean value: 2nd years			**63.1%**	
3rd year students, Module 3.5				
2001	144	28	73.5%	(2.0%)
Mean value: 3rd years			**73.5%**	
4th year students, Module 4.2				
2002	120	28	71.9%	(2.3%)
2003	100	30	70.4%	(2.5%)
Mean value: 4th years			**71.2%**	
Overall mean: 7 EOM papers			**67.5%**	
Overall mean: 15 examinations			**66.7% ± 3.9%**	

Section A: True/False items

Stratification of the data in the two types of examinations (EOS and EOM) is shown in Tables 1 and 2. The overall mean difficulty index (p-value) was 66.7% ± 3.9%, ranging from 60.4% to 73.5%. For the eight EOS examinations the mean p = 66.1%, and for the seven EOM examinations the mean p = 67.5%. No significant difference was found between the mean p-values of the EOS and EOM examinations (two sample t-test, p = 0.499 and Wilcoxon rank sum test, p = 0.464). Likewise no significant difference (two sample t-test, p = 0.473 and Wilcoxon rank sum test, p = 0.820) was found between the mean discrimination index (d-values) of the EOS (mean d = 0.22) and EOM (mean d = 0.24) examinations.

Table 2:
Discrimination Indices (d) of T/F items in EOS and EOM examinations

Students and Examinations			Discrimination Index	
Year	Number of T/F items per paper	Number of students	Mean d per paper	(SD) per paper
1st year students				
2001	140	40	0.24	(0.02)
2002	160	54	0.22	(0.02)
Mean value: 1st years			**0.23**	
2nd year students				
2001	160	30	0.25	(0.02)
2002	160	40	0.25	(0.02)
Mean value: 2nd years			**0.25**	
3rd year students				
2001	160	28	0.22	(0.02)
2002	160	30	0.18	(0.02)
Mean value: 3rd years			**0.20**	
4th year students				
2002	160	28	0.16	(0.02)
2003	144	29	0.26	(0.02)
Mean value: 4th years			**0.21**	
Mean of 8 EOS examinations			**0.22**	
1st year students, Module 1.6				
2001	140	40	0.23	(0.02)
2002	120	54	0.24	(0.02)
Mean value 1st years:			**0.23**	
2nd year students, Module 2.3				
2001	100	30	0.28	(0.02)
2002	100	40	0.28	(0.02)
Mean value 2nd years:			**0.28**	
3rd Year EOM 3.5				
2001	144	28	0.21	(0.02)
Mean value 3rd years:			**0.21**	
4th year students EOM 4.2				
2002	120	28	0.20	(0.02)
2003	100	30	0.21	(0.02)
Mean value 4th years			**0.20**	
Mean of 7 EOM examinations			**0.24**	
Overall mean of 15 examinations			**0.23**	**Mean individual items 0.22 (0.04)**

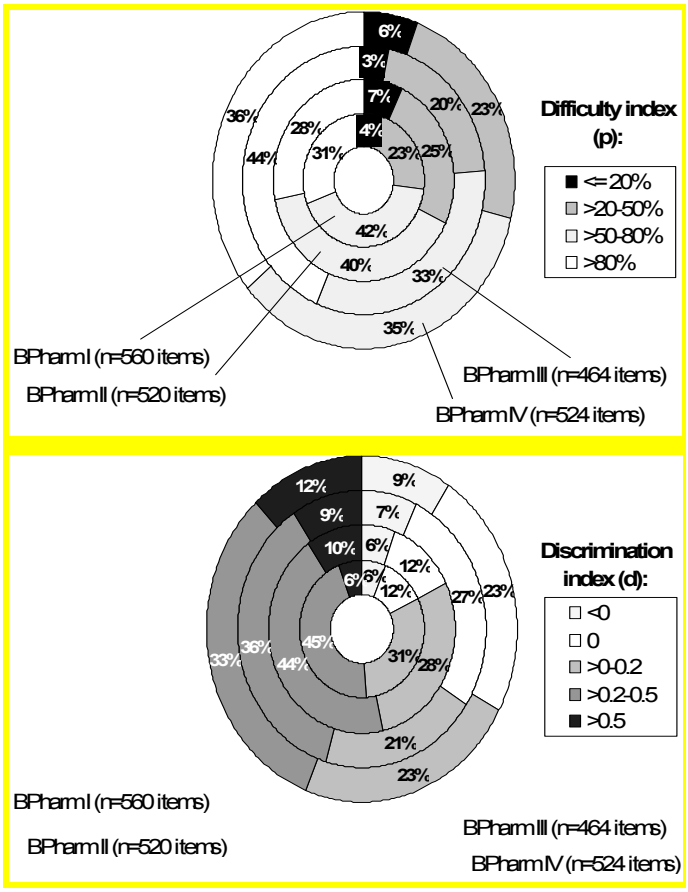
After stratification of the data in study years, the mean values for the difficulty indices of the T/F items in the examinations written by the first, second, third and fourth year students were respectively 66.2%, 63.8%, 70.9% and 67.2%. With the exception of one, no significant difference was found between the p-values of these four study years. The exception was the third- year group's examinations, which had a significant higher difficulty index than the second year group (ANOVA, $p < 0.05$). The mean difficulty indices of the examinations written in two calendar years, 2001 and 2002, were 65.4% and 68.1% respectively. These mean values of p did not differ significantly, either in the ANOVA test ($p > 0.050$) or in the two sample t-test ($p = 0.225$).

Stratification of the T/F data in study years, shows mean discrimination values of 0.23, 0.27, 0.20 and 0.21 for examinations of the respective first, second, third, and fourth year students. The only significant, but small, difference was the slightly better d-value found in the second year, compared to the third and fourth year examinations (ANOVA, $p < 0.05$).

The mean d-values of the examinations written in two calendar years, 2001 and 2002, were respectively 0.24 and 0.22. No significant difference (ANOVA, $p > 0.05$ and two sample t-test, $p = 0.334$) was found.

Analysis of the total number of 2068 single T/F items from Section A revealed that the distribution of both difficulty and discrimination indices of the specific items covered a wide range in all study years (see Figure 1). Analysis of the difficulty indices of all the examinations showed that 710 items (34%) were very easy with $p > 80\%$ (a higher % = a less difficult question), 1252 items (61%) had a moderate difficulty with $p > 20\%$ to 80% and 106 items (5%) were very difficult with $p \leq 20\%$ (see Figure 1). The distribution of discrimination indices in Figure 1 shows that 191 items (9%) discriminated well between high and low achievers ($d > 0.5$); 816 items (40%) had moderate discrimination ($d > 0.2$ - 0.5); 539 items (26%) had a very limited discrimination ($d > 0$ to < 0.2); 375 (18%) had $d = 0$, and only 147 items (7%) had negative discrimination indices. Difficulty and discrimination indices did not correlate ($R^2 = 0.06$).

Figure 1: Numbers of items in different categories of difficulty and discrimination (By year group)

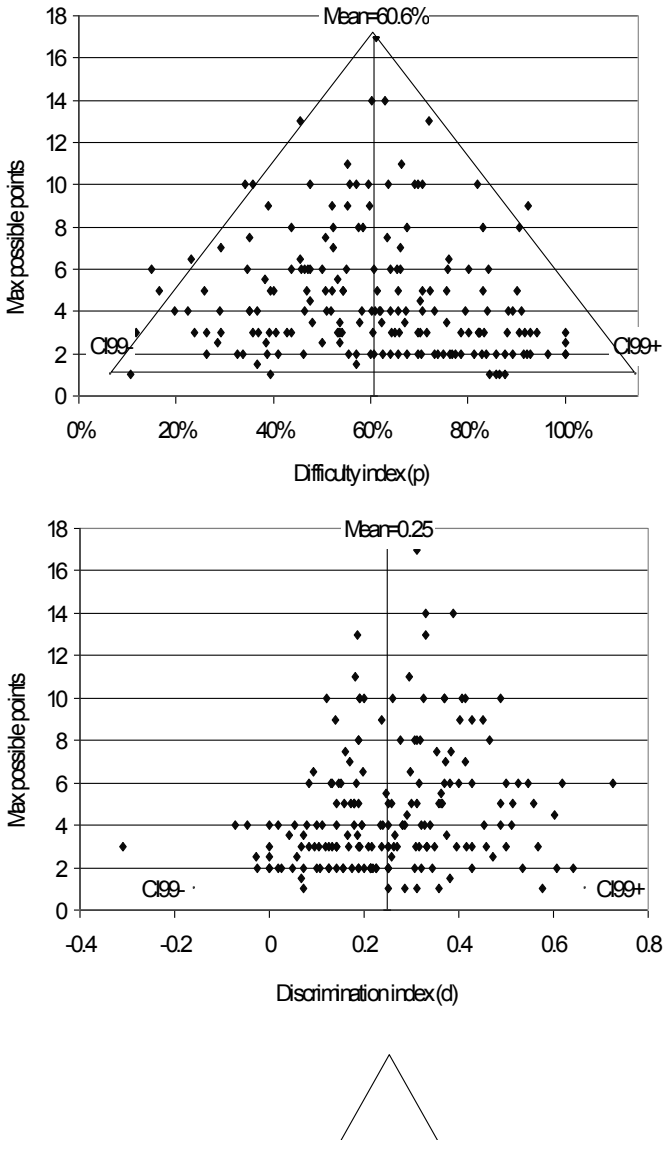


Section B: Constructed response questions

A total of 188 constructed response questions from the 15 examinations were analysed. Maximum possible marks per question ranged from 1 to 17, with a median of four marks per

question. Figure 2 illustrates difficulty and discrimination indices for questions that counted different numbers of marks.

Figure 2: Difficulty and discrimination indices of constructed response questions (n=188)

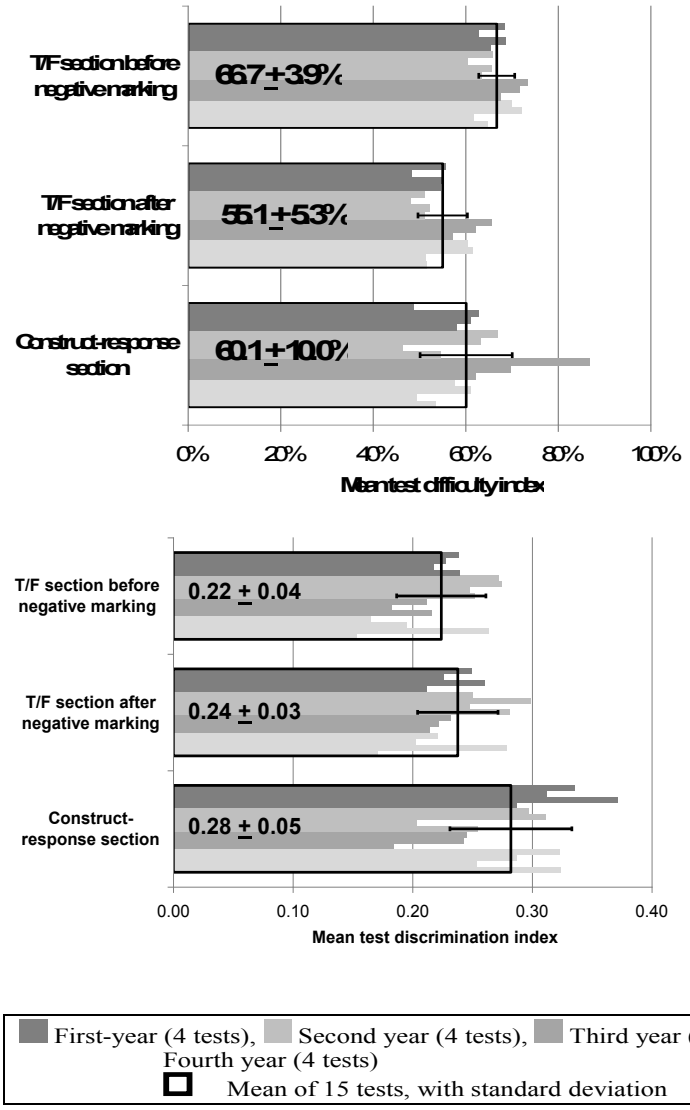


The difficulty and discrimination indices were widely scattered for the questions worth few marks, similarly to the indices found for True/False items. Questions which were worth more marks had their indices grouped more closely together so that the points on the scatter plot roughly fell into a pyramid shape with the mean value at its peak and the 99% confidence interval at the base. This finding is logical, since each mark could be considered to cover a separate item. Diverse difficulty and discrimination indices then tended to mean out within questions, which could each be regarded as a miniature "test". Difficulty and discrimination indices did not correlate ($R^2 = 0.03$).

Mean test difficulty and discrimination

Figure 3 shows the mean difficulty and discrimination indices for Section A (with and without negative marking) and Section B of each of the 15 tests analysed, with overall mean for each section type.

Figure 3: Mean test difficulty and discrimination indices (n=15) per test section, and overall mean



Negative marking significantly decreased the difficulty indices of Section A (True/False items): the mean p-values significantly decreased from 66.7% to 55.1% on average (paired t-test, $p < 0.001$). Due to negative marking, the mean discrimination indices Section A significantly increased from 0.22 to 0.24 on average (paired t-test, $p = 0.034$). Mean d-values for the True/False sections with negative marking were below 0.2 in one examination, and between 0.2 and 0.3 in the remaining 14. For constructed response sections, the d-values were below 0.2 in one examination, between 0.2 and 0.3 in eight, and above 0.3 in six examinations.

On average, the difficulty index of the constructed response sections were significantly higher (paired t-test, $p = 0.035$) than the negatively marked True/False sections (mean p = 60.1% vs 55.1%,), while their effectiveness of discrimination

was significantly higher (0.28 vs 0.24, paired t-test, p = 0.013).

Limitations of item analysis techniques

Item analysis techniques have a limited value on its own to prove the quality of assessments.

The difficulty index measures only the difficulty, expressed as the number of questions answered correctly by a percentage of students in a class. It does not measure the depth of knowledge as classified by Bloom's Taxonomy for educational outcomes in the cognitive domain (Bloom *et al,* 1956). Difficulty can be measured with electronic counting methods, but evaluation of the depth of knowledge needs an experienced examiner who reads the question with insight. Difficulty and depth of knowledge are not necessarily related, for example a difficult question may not require high order thinking, but merely good recall of detailed and difficult facts. Therefore the depth of knowledge was independently measured in the main study (Fourie, 2004). Comparison of the overall difficulty of examinations in the present study with the depth of knowledge in corresponding examinations, did not show a direct relationship between these two control parameters in the BPharm examinations studied (Fourie, 2004).

Item analysis techniques require electronic marking as standardised procedure for regular use. The manual method applied by us was time consuming.

Without detailed analysis of the single items by the person who set the test, the full benefit of the item analysis is not accomplished.

Item analysis results cannot be generalised, they are situated in a specific context, such as the particular test, level of study or academic class who wrote the examination.

Discussion

Difficulty

Table I shows consistency in the mean difficulty indices of T/F items over the four study years. The mean p-values have a narrow range (60% - 73%) and with one exception, no significant difference is found between the p-value of these examinations. The p-values are consistent from one calendar year to another - no significant difference between the 2001 and 2002 examinations. The p-values are consistent from one type of examination to the other - no significant difference is found between the EOS and EOM examinations.

The desired mean difficulty index of a test depends on its purpose. For achievement tests an intermediate mean difficulty is considered to be appropriate (Aikin, 1982). The ranges recommended by educators however vary: Professional Testing (2008) suggests 40%-90%; Kehoe (1995) recommends 30%-80%; Singh *et al.* (2003) and Carneson *et al.* (2001) 30%-70%; Ebel & Frisbie (1991), Dreckmeyer & Fraser (1991) and Streiner & Norman (1995) 20-70%. In our examinations, both the mean difficulty index for True/False items - before negative marking (p = 67%) or after negative marking (p = 55%) – are within these recommended ranges (see Figure 3).

Although the mean p-values of the T/F items are within the recommended ranges, distribution analysis (see Figure 1) indicates that 34% of T/F items are very easy (p > 80%). One could reason that this percentage is high, but in an OBE programme, certain core knowledge is required to be answered correctly by all students to prove competency. For criterion referenced tests with their emphasis on mastery-testing, many items will have p-values of 90% or above (Professional Testing, 2008). Such items will increase the difficulty index of the examination and result in questions more easily answered by the students.

The difficulty index of examinations of the third year group included in this evaluation is very high (had more easy questions compared to the other year groups): Table I shows that the mean difficulty value of the T/F section of all third-year examinations is the highest (70.9%) and significantly higher (ANOVA, P < 0,05) than those of the second years (63.8%); Figure 1 indicates that the third-year examinations have the most (44%) of the very easy (p > 80%) T/F questions; and Figure 3 shows that the construct response questions of the third-years have a higher than average difficulty index while their power of discrimination is below average. These findings lead to recommendations for the identification and improvement of third year EOS and EOM examinations by reviewing and evaluating the questions. Similar questions to those in the evaluated examinations with a difficulty index above 80% and discrimination index below 0.02 should be avoided in future. Such questions should be amended to decrease the difficulty index, and increase the discrimination index before including them in an item bank.

Comparison of the second-year examinations with the other academic years, demonstrates that the mean difficulty indices of the T/F items (with or without negative marking) are the lowest (most difficult) of all years - although only the difference between the second and third years is significant (ANOVA, p< 0.05). The second-years' questions have a small, but significantly better discrimination ability (ANOVA, p< 0.05) than those of the two more senior years. These findings are in line with the lower pass rate of the second years (see Appendix A). The lower second year pass rate may be explained by the nature and content of the second year modules at with complex clinical themes such as: Cardiac pharmacy, Respiratory pharmacy, pharmacokinetics and pharmacodynamics.

Discrimination

Overall the same extent of consistency was found with the mean discrimination values as with the difficulty values. The d-values are consistent between the study years, from one calendar year to another and between the EOS and EOM examinations.

Most educators consider 0.15 to 0.2 as the lowest d-value which still has an acceptable discrimination power (Kehoe, 1995; Singh *et al.,* 2003) and many education authorities encourage test writers to aspire for discrimination indices

beyond 0.3 (Dreckmeyer & Fraser, 1991). Negative values are generally considered to be undesirable (Singh *et al.,* 2003). Questions with a perfect discrimination are not often found (Linn & Gronlund, 2000).

This study establishes mean discrimination indices above 0.2 for T/F questions in Section A and B (see Figure 3). Although these mean d-values are low, they are within the acceptable range. Forty nine percent of the T/F items and 59% of constructed response questions display acceptable discrimination power (d> 0.2). The consistently low discrimination indices of all examinations and the large number of T/F questions with a d-value smaller than 0.2, leave room for improvement. The items with a negative d-value should not be included in item banks. The large number of items with a low d-value may be partly due to the high percentages of easy items in our examinations, as items which are answered correctly by a large proportion of examinees have a markedly reduced power to discriminate (Kehoe, 1995; Sim & Rashiah, 2006). Sim and Rasiah (2006), found similar results to ours in para-clinical multidisciplinary papers of medical students: on average 38% of their True/False type MCQs were very easy and two-thirds of those very easy items had very poor or negative discrimination.

The constructed response sections of all first-year examinations have an above-average effectiveness of discrimination (see Figure 3). This finding could be attributed to the diverse educational experiences and prior knowledge of the first years. The small group educational approach used in the problem-based learning method of the BPharm programme may provide the low achievers with opportunities to develop their abilities in successive years and hence close the performance gap between themselves and their more advantaged classmates. This may be one possible reason why questions from second year onwards do not discriminate as well between the students.

This study found that the discrimination indices of single items vary greatly. A factor which might have contributed to this variation is the diversity of the learning content covered by each summative examination. Each module integrates different disciplines and each test item measures a different aspect of the complex learning content. A student's ability to perform well in a single question will not necessarily reveal his ability as judged by his overall examination mark.

Linn and Gronlund (2000) alert us that the most important question in assessment is whether an item measures a core learning outcome, not how high the discrimination power of the item is. Learning outcomes for the modules of this pharmacy programme were derived from the draft competency standards for entry-level pharmacists (Summers *et al*, 2001; Interim Pharmacy Council of South Africa, 1998), developed by the South African Pharmacy Council, the national authority for pharmacy education. Content validity of these examinations was established in the larger study by matching on a grid the questions in each EOM examination with the learning objectives for the corresponding module (Fourie, 2004).

Conclusions

Compared to standards suggested in the literature, the mean difficulty and discrimination index values of the 15 examinations in the study are within the recommended ranges. The high degree of consistency in these values establishes longitudinal and vertical coherence and contributes to the reliability of these assessments. These findings point out, that the level of difficulty increased with the years relative to the knowledge and capability of the students. The acceptable overall level of difficulty and discrimination, as well as the consistency in these values, are considered as attributes to the quality of the examinations.

Due to the negative marking applied as the standard practice in the BPharm programme, the students performed better in the True/False than in the constructed response questions. This finding does not imply that the True/False questions *per se* were more difficult than the constructed response questions, as the analysis of the True/False questions without negative marking proved the contrary. Authors like Sim and Rashiah (2006), Downing (2002) as well as Dent and Harden, (2001) express the view that True/False (selected response) questions likewise require less depth of knowledge or higher level thinking than constructed response questions.

The calculation method, which we adapted, is useful for the analysis of questions counting more than one mark (such as constructed response questions).

The main focus of item analysis is on the performance of the separate questions. The largest benefit to be gained is that it allows each examiner to determine whether the items he wrote functioned in the way it was intended to do. Reviewing the separate items and the students' responses to it, provide information as to whether the item measures at the correct level of difficulty for the specific test or examination and whether it distinguishes those who know the content from those who do not (McCown *et al.,* 1996). Appropriate questions can be included in an item bank (Rudner, 1998). Questions which should get special attention are very easy questions (p> 80%), include them in the item bank only if they contain essential/core knowledge. Very difficult questions (p< 20%) should be screened for inaccuracy, irrelevant or inappropriate content, misleading terms; unnecessary or difficult language. Items with negative discrimination should be screened for flaws and amended or discarded. Hopefully this article will encourage others to examine their test-writing, and share similar ideas and best practices for developing testing materials for pharmacy courses.

Acknowledgements

The authors wish to thank the National Research Foundation, Social Sciences and Humanities, for financial assistance (NRF Grant No 15/1/3/20/0004), Dr Penelope Richards for introducing us to the concepts of difficulty and discrimination indices, Prof Coryce Haavik for her interest, contributions and

support; Prof HS Schoeman for statistical analysis; Ms M Lategan for data capturing; the staff of the Pharmacy Department for assessing the examinations.

References

Aiken, L. R., (1982). Psychological testing and assessment. 4th Ed. Boston, MA: Allyn and Bacon.

Anderson, H. M., Anaya, G., Bird, E. & Moore, D. L. (2005). A review of educational assessment. American Journal of Pharmaceutical Education, 69(1), 84-100.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: the classification of educational goals. Handbook I, The cognitive domain. London: Longman Group.

Carneson, J., Delpierre, G. & Masters, K. (1996). UCT's Designing and managing multiple choice questions. University of Cape Town, South Africa. Online available from: http://web.uct.ac.za/projects/cbe/mcqman/mcqman01.html. Accessed January 31, 2001.

Dent, J. A. & Harden, R. M. (2001). A practical guide for medical teachers. Edinburgh: Churchill Livingstone.

Downing, S. M. (2002). Assessment of knowledge with written test forms. In G. R. Norman, C. P. M. van der Vleuten, & D. I. Newble (Eds.), International handbook of research in medical education, Part Two, Vol. 7 (pp. 647-672). Dordrecht: Kluver Academic Publishers.

Dreckmeyer, M. & Fraser, W. J. (1991). Classroom testing in biology and physical science. Pretoria: HAUM.

Ebel, R. L. & Frisbie, D. A. (1991). Essentials of educational measurement, 5th Edn. Englewood Cliffs: Prentice-Hall.

Fourie, S. (2004). Knowledge assessment in the MEDUNSA BPharm degree programme. [Doctoral thesis]. Medunsa: Medical University of Southern Africa.

Interim Pharmacy Council of South Africa. (1998). Draft unit standards for future pharmacists at entry level, Pretoria: South African Pharmacy Council.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. Practical Assessment, Research & Evaluation, 4(10). ERIC/AE Digest. ERIC Document Reproduction Service ED398237. Online available from: http://PAREonline.net/getvn.asp?v=4&n=10. Accessed March 12, 2003.

Linn, R. L. & Gronlund, N. E. (2000). Measurement and assessment in teaching. 8th Edn. Upper Saddle River, NJ: Prentice-Hall.

Mabope, L. A. & Meyer, J. C. (2009). Ten year's experience: access to and throughput of the BPharm programme at the University of Limpopo, Medunsa Campus. Paper presented at the Medunsa Academic Day, 13 August 2009. Medunsa.

McCown, R.R., Driscoll, M., & Roop, P. G. (1996). Educational psychology: a learning-centred approach to classroom practice. 2nd Edn. Boston, MA: Allyn & Bacon.

Professional Testing. (2008). Building high quality examination programmes. Online available from: http://www.proftesting.com/test_topics/steps.php Accessed January 14, 2010.

Rudner, L.M. (1998). Item Banking. Practical Assessment, Research & Evaluation 6(4). Online available from ERIC Clearinghouse on Assessment and Evaluation: http://ericae.net/pare/getvn.asp?v=6&n=4. Accessed August 2, 2002 and http://PAREonline.net/getvn.asp?v=6&n=4 accessed February 10, 2010.

Sim, D. S-M. & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in True/False-type multiple choice questions of a para-clinical multidisciplinary paper. Annals of the Academy of Medicine, Singapore, 35:67-71.

Singh, T., Singh, D., & Vinod, K. P. (2003). Better medical education: principles of medical education. Indian Acadamy of Pediatrics. Online available from: http://www.edu4med.com/princpl/index.html. Accessed June 24, 2003.

Streiner, D. L. & Norman, G. R. (1995). Health measurement scales: a practical guide to their development and use. 2nd Edn. Oxford: Oxford University Press.

Summers, R., Haavik, C., Summers, B., Moola, F., Lowes, M., & Enslin, G. (2001). Pharmaceutical education in the South African multicultural society, American Journal of Pharmaceutical Education, 65(2), 150-154.

Appendix 1: BPharm Pass rates (Mabope & Meyer, 2009)

BPharm pass rates 1999-2003

Year	BPharm 1		BPharm 2	
	Exam/ Passed	%	Exam/ Passed	%
1999	30/29	96.7		
2000	36/32	88.9	29/29	100
2001	46/43	93.5	32/30	93.8
2002	57/54	94.7	44/37	84.1
2003	40/40	100	60/56	93.3
Ave		**94.8**		**92.8**

BPharm 3		BPharm 4		Annual Mean %
Exam/ Passed	%	Exam/ Passed	%	
				96.7
				94.5
29/28	96.6			**94.6**
31/30	96.7	28/28	100	**93.9**
37/36	97.3	30/30	100	**97.7**
	96.9		**100**	**95.5**

Students must pass wide variety of continuous and summative assessments during the year as well as the final examinations at the end of each of 6 modules (EOM) and two semesters (EOS) per year. The University pass mark is 50%.

Appendix 2: Examples of True/False and Constructed response questions

Section A: Contained True/False items, grouped into sets of four.
In the examinations negative marking is applied. Each correct response counts 0.5 marks, each wrong response counts ‑ 0.5 marks, but no set of four items can score less than zero.

Example:
South African guidelines for drugs used in the treatment of tuberculosis in pregnancy include the following points:
A Rifampicin is contraindicated (F)
B Isoniazid is used with pyridoxine supplementation (T)
C Steptomycin is not advisable to be used (T)
D Ethambutol is used, provided that the dose should not exceed 15 mg/kg/ (T)

Section B: Contained constructed response questions counting 2 to 18 marks each.

Example (question for 4 marks):
A 70 kg male with COPD receives the normal dose of 1200 mg theophylline per day. His mean steady state plasma level is 22.5 mg/l. Calculate the theophylline clearance in l/kg/hr for the patient. Assume a bioavailability of 1 and salt factor of 1. (3)
Comment on the clearance of the patient (normal clearance is 0.04 l/kg/hr) (1)