MOTIVATING TEACHERS BY ASSIGNING EACH A DIFFERENT GROUP OF MEDICAL STUDENTS TO EDUCATE ON A SAME PHYSIOLOGY CURRICULUM

MIHAIL GHEORGHE GLIGA, MARIUS SABĂU

Physiology Department, University of Medicine and Pharmacy Tîrgu-Mureş

Abstract

Aim. *We intended to prove that if we construct groups of students by systematically sampling them from the whole alphabetically sorted group, then education on a same physiology curriculum of each different group by a different teacher can be a better alternative to the actual situation: one curriculum – one teacher. We will further name these groups 'Alpha Groups' (AGs).*

Material and Methods. *We used 4 large AGs of about 160 students and 4 small AGs of 40 students who were tested, during our University 2012 admission session, on a human biology (pre-physiology) curricula by 100 Multiple Choice Questions (MCQs) of which 75 questions were evaluating mainly memorizing skills and 25 MCQs were mainly thinking questions. We calculated the average and variance of marks for each AG on each type of MCQ and we analyzed for each AG the values of P and D indexes in each MCQ. We statistically compared the results between AGs.*

Results. *We found no significant differences, even between small AGs of 40 students, for none of the parameters. Therefore, if AGs would be evenly educated on a same physiology curriculum by different teachers, they should have the same results on a same neutral final evaluation. We could link MCQ quality to the teachers.*

Conclusions. *If AGs have significantly different results on a proper MCQ test, that will probably be due to the unequal education they received and this should motivate each teacher to educate well his group of students. Furthermore, analysis of MCQs can motivate teachers to write good quality MCQs.*

Keywords: motivating teachers, medical education, physiology curriculum, cognitive domain, MCQ analysis.

MOTIVAREA PROFESORILOR PRIN ATRIBUIREA FIECĂRUIA A UNUI GRUP DIFERIT DE STUDENŢI MEDICINIŞTI PENTRU A-I EDUCA PE BAZA UNUI ACELAŞI CURICULUM DE FIZIOLOGIE

Rezumat

Obiective. *Am dorit să arătăm că, dacă alcătuim grupuri de studenţi prin eşantionarea sistematică a unui întreg grup în care studenţii au fost ordonaţi alfabetic în prealabil, atunci educarea, pe un acelaşi curriculum de fiziologie, a fiecărui grup de câte un profesor diferit poate fi o alternativă mai bună la situaţia actuală: un curriculum – un profesor. Vom numi în continuare aceste grupuri alfagrupuri.*

Materiale şi Metode. *Am folosit 4 alfagrupuri mari de aproximativ 160 de studenţi şi 4 alfagrupuri mici de 40 de studenţi care au fost testaţi, în timpul sesiunii de admitere 2012 a universităţii noastre, din curriculumul de Biologie Umană (pe care îl considerăm ca un curriculum pre-fiziologie) prin 100 de întrebări cu răspunsuri multiple (MCQs), dintre care 75 de întrebări au evaluat în special capacitatea de memorare, iar 25 de MCQs au fost în principal de gândire. Am calculat parametri ca media şi varianţa notelor pentru fiecare grup şi pe fiecare tipuri de întrebări şi am analizat, pentru fiecare grup, indicii P şi D pentru fiecare MCQ în parte. Am comparat statistic rezultatele între alfagrupuri.*

Rezultate. *Nu am găsit diferenţe semnificative nici măcar între alfagrupurile mici de 40 de studenţi pentru nici unul din parametrii studiaţi. În consecinţă, dacă alfagrupurile vor fi educate în egală măsură pe un acelaşi curriculum de fiziologie de către profesori diferiţi, atunci ele vor avea aceleaşi rezultate la o aceeaşi evaluare finală neutră. Am putut face o legătură între calitatea unui MCQ şi profesori.*

Concluzii. *Dacă alfagrupurile de studenţi vor avea rezultate semnificativ diferite la un test MCQ corespunzător calitativ, aceasta se va datora probabil educaţiei inegale pe care au primit-o şi acest lucru ar trebui să motiveze fiecare profesor să educe bine grupul său de studenţi. În plus, analiza MCQ poate să motiveze profesorii să scrie MCQ de bună calitate.*

Cuvinte cheie: motivarea profesorilor, educaţie medicală, curriculum de fiziologie, domeniul cognitiv, analiza MCQ.

INTRODUCTION

Education on a given medical curriculum, for example the physiology curriculum, is provided to students in our university in all 3 domains: cognitive, by lectured courses, affective, mainly by teacher-student interaction and psychomotor by laboratory classes [1-4]. Until now, students' knowledge in the cognitive domain has been evaluated mainly by final semester oral exams. The marks students obtained in these exams are the only ones that rank them later and indicate the education efficacy. Since, for a given medical physiology curricula, we have a single professor who teaches and also evaluates students' knowledge in the cognitive domain at the end of their course, some inconsistencies may consequently occur in tracing teachers' education efficacy [5-9]. For example if the students' exam results are poor we cannot distinguish whether the students' learning capacities are poor or the education they received was not sufficient quantitatively or qualitatively. Moreover, if the leadership of the University wants to motivate teachers to perform better in order to increase education efficacy on a given curriculum, the process cannot be controlled [10]. This means that the leadership of the University cannot distinguish if better students' exam results are the consequence of better education or simply the consequence of fewer expectations in evaluating students by the same teacher who educated them. Within this paper we aim to search for alternatives to this situation.

First we intend to demonstrate that, if we configure groups of students by systematic sampling [11] them from the whole alphabetically sorted group, there are no significant differences in learning capacities between those groups or at least that they have the same memorizing and thinking capacities. For example if we want to obtain such 4 groups of students we have to sort all of them alphabetically and then every fourth student will go into a group. We also intend to find out what is the minimum number of students in AGs down to which our theory is valid.

Second, if it is decided that each different group of students will be educated by a different teacher, then a neutral and identical evaluation of all students would be necessary. This can be done by securitized protocols in many written formats [12,13] from which one of the most common is by MCQ testing. We wanted to show that choosing the MCQ format tests can lead to a good evaluation of students because the quality of the MCQ set used for tests can be analyzed. Analysis of the quality of an MCQ set is done first by calculating for each MCQ the item difficulty P index and item discrimination D index. Item P index is the ratio between the number of correct responses to the total number of responses for that test item, and item discrimination D index is, given a group of students ranked by their scores on the MCQ set test, the difference between the P index of the top 27% scorers minus the P index of the bottom 27% scorers [14]. Consequently P index can have values between 0 and 1 and the bigger it is, the easier that MCQ is, and the D index can have values between -1 and +1. An MCQ with a good discrimination quality has a D index between 0.3 and 0.4 while a very good MCQ has a D index over 0.4 and, on the opposite, MCQ with D indexes between 0.2 and 0.3 are marginal in quality while those with a D index of less than 0.2 are poor and should be revised or eliminated from the set [15]. Usually, good and very good qualities, according to a discrimination D index are obtained by an MCQ with a difficulty P index between 0.25 and 0.75 or even restricted interval 0.4 to 0.6 [16].

MATERIAL AND METHODS

We used 4 large AGs obtained from 621 candidates for Medicine (M group) and we named them as follows: M1=159 candidates, M2=151 candidates, M3=159 candidates, M4=152 candidates. Small differences between M groups are due to candidates who did not show up at the exam. We also used 4 smaller AGs obtained from 163 candidates for Dental Medicine (S group) and we named them S1=41 candidates, S2=41 candidates, S3=41 candidates and S4=40 candidates. AGs were obtained by the intrinsic protocol of the exam for admission in our University in year 2012. All AGs from 1 to 4, either M type or S type, were evaluated on the same 100 MCQs that we will further

Manuscript received: 24.09.2012
Accepted: 07.10.2012
Adresses for correspondence: mihaigliga@yahoo.com

name Base100MCQ, but randomly mixed to generate 1 to 4 variants of different ordered 100 MCQs. Each MCQ consisted of a statement with 5 answers, identified A to E of which 1 or 2 could be correct, and the evaluation of each MCQ was in an all or nothing way. The MCQ set content covered a Human Biology Curriculum which we agree to be a pre-Physiology Curriculum. The Base100MCQ consisted of two distinct types of questions. The first one, named Old 1-75 MCQs, contained MCQs numbered 1 to 75 which were randomly computerized and extracted from a 1430 MCQs set previously published together with the correct answers some years ago. The second type of questions, named New 76-100 MCQs, were constructed just before the exam and therefore were unknown to candidates. We assumed that Old MCQs were mainly memorized by candidates and that New MCQs asked for more thinking than memorizing.

Since the currently employed computerized protocol for the evaluation of the admission exam results of candidates did not provide us with the information required in order to compare AGs between them and on each type of MCQ, we had to use raw data obtained by optical computerized evaluation of students' answering papers. These raw results for each candidate consist of an identification number badge, to which a set of data are assigned such as name, baccalaureate mark, language in which he desired to perform the exam, faculty to which he applied (medicine or dental medicine), then the MCQ variant he was given at the exam and finally a 5x100 sequence of 0 and 1 which represent the candidate's answers. A 0 is employed where the answer is considered to be false and a 1 where the answer is marked as true. For example candidate 2308 answered the third question, in his numbered 4 variant, by marking the 'C' answer to be correct:

2308 4 0000100001001001000000001001000000010
0010000100100001100000100001000000100101010000001
0000101001010000001000010011000101010100101001 10
000010001000000101010011000101000000100001010001 1
0000001010001000100010001001000100000100010010 00
1010010011000110000000100001001100000010000101010
0000100011001000101001001011001000110100011001 0
0100101001000000110001100010010010000010001100010
0100101000000010100000100011011000010010100000010
1100001100001001010110001100000010100010100001 00
01000100010000001110010000100110000010101010

We evaluated by comparison the correct set of 5x100 zeroes and the set given ineach candidate's answer for each MCQ and we marked each of the (621+163)*100 = 78400 MCQs with True, if the answer was correct to that MCQ or False, correspondingly. This was necessary in order to be able to correctly count answered MCQs out of total answered MCQs by students in an AG and thus to be able to calculate the P index for each MCQ in each AG. We ranked students in each AG by their number of correct/True MCQs out of 100 so we could calculate the D index of each MCQ for each AG. We further rearranged the order of MCQs from all the variants to obtain same Base100MCQ configuration for each AG. This was necessary in order to be able to compare D versus P correlation between groups. We calculated each student's mark and we checked our whole work until present by comparing it with the official results. We calculated students' partial marks as if they were evaluated only on Old MCQ or New MCQ. We finally compared AGs by the overall admission statistics results and by the average and variation of the students' total or partial marks.

RESULTS

Overall statistic results of AGs were compared by observed distribution frequency test Chi^2 as they are shown in Table I. There are no significant differences between large Medicine M1, M2, M3, M4 AGs or between smaller Dental medicine S1, S2, S3, S4 AGs, neither concerning admitted versus rejected candidates nor concerning admitted without tax / budget candidates versus admitted with tax candidates.

We ranked all the 163 candidates for Dental medicine by the total 100 MCQs marks, which are the same marks as those retrieved from the official results after evaluation of the biology exam in the 2012 admission session of our University. In Figure 1, we showed separately the classification inside each AG according to our methods. Moreover, the calculated partial marks for Old MCQs and New MCQs are also shown for each candidate and it is obvious that candidates were more successful in each AG on Old, probably memorized MCQs, than on New MCQs, which were unknown to them up to the moment of the exam and for which they had to think more. Notice that the trend line for Old MCQ Marks is smoother than the trend line for New MCQ Marks, showing that two consecutive overall biology ranked students have similar memorizing skills but may have different thinking skills and this characteristic is the same inside all 4 small AGs. We found that P index

Table I. The overall statistics of AGs results after our University 2012 admission biology exam.

M	Admit.	Reject.	**M**	Budget	Tax	**S**	Admit.	Reject.	**S**	Budget	Tax
M1	78	81	**M1**	58	20	**S1**	25	16	**S1**	18	7
M2	64	87	**M2**	49	15	**S2**	22	19	**S2**	15	7
M3	82	77	**M3**	62	20	**S3**	20	21	**S3**	14	6
M4	77	75	**M4**	57	20	**S4**	24	16	**S4**	15	9
χ^2	**p = 0.367**		**χ^2**	**p = 0.984**		**χ^2**	**p = 0.653**		**χ^2**	**p = 0.907**	

distribution presents, in each AG, right skewness for Old MCQs, meaning that those questions were easy for most candidates in each AG, and left distribution for New MCQ, meaning these questions were difficult for most candidates in each AG. Similar results were obtained for the 621 candidates for Medicine and corresponding 4 large M AGs.

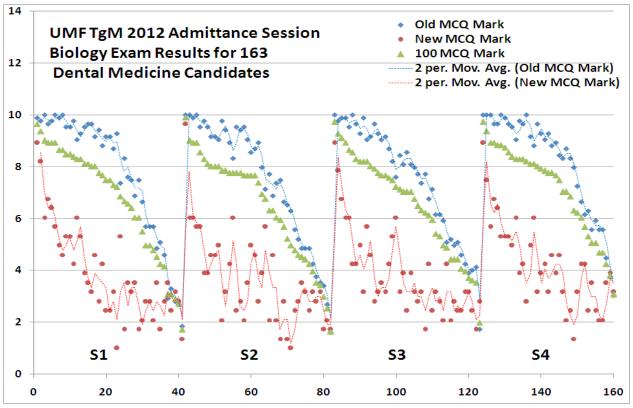
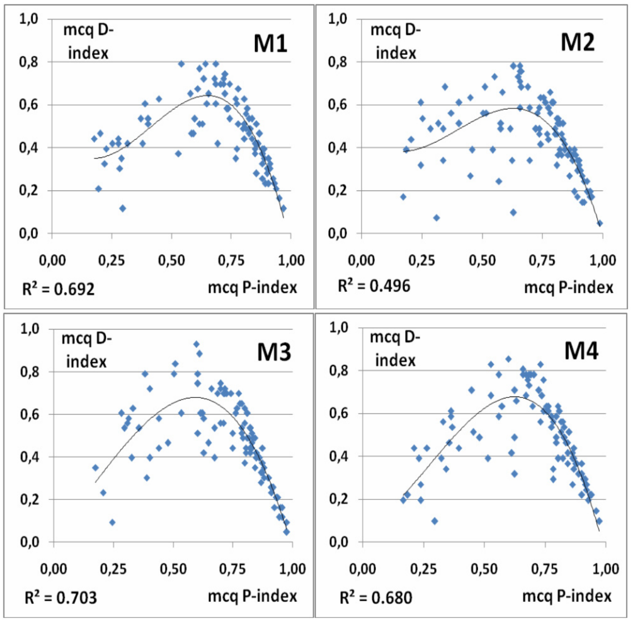


Figure 1. Comparison between AGs of partials and total marks at the biology admission session exam.

A more precise evaluation for finding significant differences between AGs is the comparison of variance and average of candidates' total and partial marks between AGs. We used for this, respectively, Fisher test and T-Student test and p values that these tests returned are shown in Table II.

An F-test returns the two-tailed probability that the variances in array1 = students' marks in AG M1 and array2 = students' marks in AG M2 for example, are not significantly different. A Student's t-Test returns the probability of being wrong when saying that average of array 1 differs significantly from average of array 2. According to the results it is unlikely that M AGs or S AGs differ between each other neither on all 100 MCQs results, nor on memorizing skills (Old MCQs) or on thinking skills (New MCQs), neither on average skills, nor on diversity of skills, because we couldn't find any $p < 0.05$ for any kind of comparison.

Correlation between D index as a function of P index of 100 MCQs for all M AGs are shown in Figures 2 to 5. Similarities suggest that difficulties of questions have the same impact in ranking students inside groups. There is a slight difference between M2 and the rest of M AGs but this aspect will be further investigated since we found the same difference in the same type of correlation for S AGs.



Figures 2, 3, 4, 5. AGs behave similarly according to D-index as a function of P-index.

There are no overall significant differences between either M 1 to 4 or S 1 to 4 AGs in the way candidates answered to poor quality questions by guessing to difficult MCQs or memorizing the easiest MCQs, or in the way good and very good MCQ discriminates them. In Table III, notice that comparison between AGs by number of

Table II. Comparison between AGs of the variance and average of candidates' total and partial marks.

Compare Marks - p values-	M1/ M2	M1/ M3	M1/ M4	M2/ M3	M2/ M4	M3/ M4	S1/ S2	S1/ S3	S1/ S4	S2/ S3	S2/ S4	S3/ S4
100 MCQ F-Test	0.409	0.768	0.961	0.592	0.387	0.734	0.779	0.467	0.801	0.654	0.979	0.637
100 MCQ T-test	0.863	0.475	0.686	0.366	0.809	0.266	0.561	0.697	0.851	0.828	0.434	0.552
Old MCQ F-Test	0.386	0.380	0.939	0.999	0.350	0.345	0.811	0.429	0.830	0.581	0.982	0.568
Old MCQ T-test	0.709	0.711	0.714	0.444	0.994	0.457	0.623	0.678	0.926	0.921	0.554	0.604
New MCQ F-Test	0.795	0.469	0.852	0.330	0.942	0.367	0.748	0.524	0.542	0.752	0.771	0.982
New MCQ T-test	0.615	0.135	0.705	0.315	0.380	0.064	0.475	0.891	0.609	0.541	0.205	0.495

Table III. Comparison of the observed frequences of quality types MCQ according to D index.

100 MCQ	Poor	Marginal	Reasonably Good	Very Good	100 MCQ	Poor	Marginal	Reasonably Good	Very Good
D - index	< 0.20	0.20-0.30	0.30-0.40	> 0.40	**D - index**	< 0.20	0.20-0.30	0.30-0.40	> 0.40
M1	3	9	21	67	**S1**	7	8	14	71
M2	12	8	23	57	**S2**	14	4	11	71
M3	8	10	16	66	**S3**	15	13	12	60
M4	6	12	18	64	**S4**	11	10	12	67
Total	29	39	78	254	Total	47	35	49	269
Average	**7.25**	**9.75**	**19.5**	**63.5**	**Average**	**11.75**	**8.75**	**12.25**	**67.25**
All M	4	13	13	70	All S	3	12	10	75
χ^2	**p = 0.415**				χ^2	**p = 0.369**			

specified quality MCQ types, using frequency distribution Chi^2 test, is a valid one since all the theoretical/expected frequencies (not shown) are greater than 5. Interesting to see that D index average of AGs can differ from D index of the whole group, from which AGs were extracted.

We also calculated for each items P index and D index according to the entire M group and S group. For the latest S group we presented the results in Figure 6, where items were ordered as in the Base100MCQ.

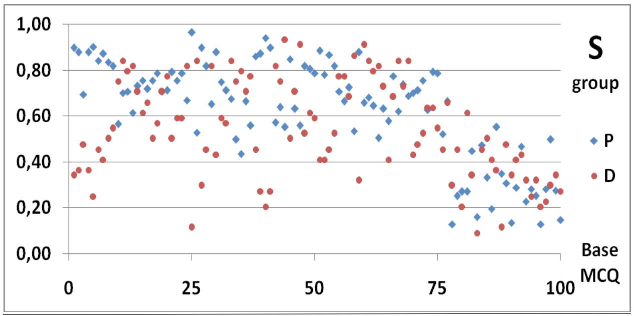


Figure 6. P index and D index of Base100MCQ according to 163 Dental Medicine candidates.

If we further sort MCQs from figure 6 by D index we obtain results shown in Figure 7. We can see that the best discrimination of Dental Medicine candidates (the S group), was obtained with MCQs having a P index between 0.4 and 0.8. We also obtained good quality MCQs, according to D index, for associated P index values between 0.2 and 0.9. Simillar results were obtained for Medicine candidates (M group).

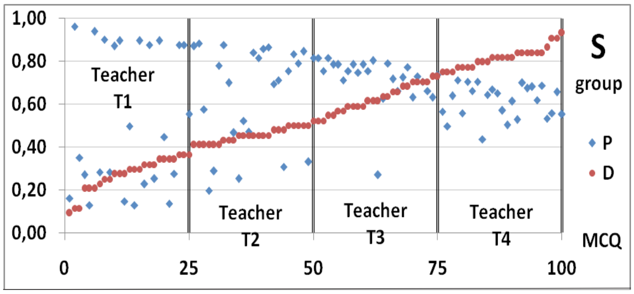


Figure 7. P index of Base100MCQ sorted on D index for 163 Dental Medicine candidates.

DISCUSSIONS

Implication of teachers in writing good quality Multiple Choice Questions is obviously very important [17,18] and can be enhanced as we will show. Evaluation of AGs in identical conditions can be readily done by MCQ format tests even during / instead the lectured courses [19]. Let us assume that, in the construction of the 100 MCQ presented in Figure 7, four teachers have participated in the following way: the first teacher T1 conceived questions 1 to 25, the second teacher T2 wrote questions 26 to 50 and so on, questions 51 to 75 were elaborated by teacher T3 and questions 76 to 100 belong to teacher T4. It is obvious that teacher T4 wrote significantly better questions than teacher T1 meaning that teachers' T4 MCQs set better differentiates between students. This hypothetical but potential hierarchy among teachers according to D index items should motivate them in writing good quality MCQs [20].

Also notice from Figure 7 that P index interval values 0.2 ÷ 0.9 which correspond to high D index values 0.4 ÷ 0.6 and P index 0.4 ÷ 0.8 corresponding to very high D index values 0.6 ÷ 0.9 are larger than usual [16]. This fact suggests that cognitive skills of the candidates are dispersed on a broad range. More, the fact that the distribution of MCQ P index inside these intervals presents a skewness to the right, as also seen in figure 7, may indicate, and not contradict Figure 1, as well as the fact that the lower half of candidates' skills range was low. Another way to enhance teachers' involvement in writing good quality material for students is by the protocol we proposed for calculation of Current P and D indexes of MCQs if they are used to teach students during the semester [21]. This computerized protocol offers teachers the possibility to refine the poor quality questions or to adjust current overall difficulty of the MCQ set according to student performances and/or teacher expectancies.

Item response theory [22] changes the way students are confronted with MCQs and how a final classification of their performances is done in order to rank them. Many people agree that the overall evaluation results are similar to those used in classical test theory, but quicker. Anyway, according to both theories, we can still evaluate AGs of students in a neutral way by MCQ tests, so educating each of those groups by a different teacher remains the same, given the same circumstances.

CONCLUSIONS

In search for alternatives to the current most frequent situation: one curriculum – one teacher, we propose a structure that can satisfy all three basic issues: good education, good assessment and traceability of these processes. In order to accomplish this structure we found that:

a. AGs of 40 students or more, constructed by systematic sampling from a whole group of students, previously alphabetically ordered, are likely to have the same skills in lower levels of the cognitive domain, such as memorizing/'recall' and 'understanding'/thinking. Since we evaluated our groups on the same pre-physiology curriculum, it is further more likely that, if these groups of students will be evenly educated on the same physiology curriculum by different teachers, they should further have the same results on a same neutral final evaluation. If AGs of students will have significantly different results in identical evaluation conditions, it will probably be due to the unequal education they received. Furthermore, if we know that each group was taught an equal number of hours, then the education of the group with poorer results was probably of poor quality teaching. Therefore the competition between AGs each educated on a same curriculum by a different teacher should motivate teachers to educate well.

b. Evaluation of AGs in identical conditions can be readily done by MCQ format tests. Quality of an MCQ set can be quantified by putting a substantial number of MCQ in a set, in order to thoroughly cover the given curriculum, and can be done by equally involving more than one teacher, for example all those teachers who educated AGs. Each teacher will write a number of questions and the post exam analysis of the quality of MCQ will show which professor conceived the best or the poorest questions and this should motivate teachers to write good and very good MCQs.

c. Finally but not least important, the number and textual content itself of a substantial MCQ set is a retrievable trace of the curriculum for which the students were assessed and also for which skill levels in the cognitive domain were assessed. And finally the most important, in the given conditions, student results will mirror the quality of education they received from their teachers.

References

1. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Pract, 2002; 41:212-264.
2. University of Oregon Teaching and Learning Center. Bloom's Taxonomy of Cognitive Levels [Internet]. [updated 2011 October 10; cited 2012 September 16]. Available from: http://tep.uoregon.edu/resources/assessment/multiplechoicequestions/blooms.html
3. Medical Education: Enhancing Learning in the Affective (Feeling) Domain. Retrieved September 2012, from: http://www.cdtl.nus.edu.sg/link/pdf/jul2003.pdf
4. Learning Taxonomy – Simpson's Psychomotor Domain. Retrieved September 2012, from: http://assessment.uconn.edu/docs/LearningTaxonomy_Psychomotor.pdf
5. Evaluate Your Own Teaching. Retrieved September 2012, from: http://www2.warwick.ac.uk/services/ldc/resource/evaluation/teaching/
6. Smain Bekhti, Nada Matta. Traceability and knowledge modelling. Retrieved September 2012, from: http://www-sop.inria.fr/acacia/WORKSHOPS/ECAI2002-OM/Actes/Bekhti.pdf
7. H. Richter, C. Miller, G. D. Abowd, and H. Funk. Tagging Knowledge Acquisition To Facilitate Knowledge Traceability. International Journal on Software Engineering and Knowledge Engineering, 2004; 14(1):3-19
8. Effective Teaching Strategies. Retrieved September 2012, from Whole Person Education: http://www.asa3.org/ASA/education/teach/methods.htm
9. Effectiveness of Problem-based Learning Curricula: Research and Theory. Retrieved September 2012, from: http://www.med.uni-frankfurt.de/lehre/fam/literatur/container_journal_club/effectiveness_Colliver_Volltext.pdf
10. Process Controllers. Retrieved September 2012, from: http://www.engineeringtoolbox.com/process-controllers-d_499.html
11. Saint Joseph's University, Department of Psychology. Retrieved September 2012, from: http://schatz.sju.edu/methods/sampling/random.html
12. Catforms Testing Service. Item Formats & Standards [Internet]. [updated unknown; cited 2012 September 16]. Available from: http://www.catforms.com/pages/Item-Formats-%26-Standards.html
13. The University of North Carolina. Improving Multiple Choice Questions [Internet]. [updated unknown; cited 2012 September 16]. Available from: http://cfe.unc.edu/pdfs/FYC8.pdf
14. Si-Mui Sim, Raja Isaiah Rasiah. Relationship Between Item Difficulty and Discrimination Indices in True/False-Type Multiple Choice Questions of a Paraclinical Multidisciplinary Paper. Ann Acad Med Singapore, 2006, 35: 67-71.
15. Item Discrimination Indices [Internet]. [updated unknown; cited 2012 September 16]. Available from: http://www.rasch.org/rmt/rmt163a.htm
16. Professional Testing [Internet]. [updated unknown; cited 2012 September 16]. Available from: http://www.proftesting.com/test_topics/steps_9.php
17. Carroll RG –Design and Evaluation of a National Set of Learning Objectives: The Medical Physiology Learning Objectives Project. Adv Physiol Educ, 2001; 25 (2):2-7.
18. International Database for Enhanced Assessments & Learning. Retrieved September 2012, from IDEAL Consortium: http://www.idealmed.org/homeindex.html
19. Silverthorn DU-Teaching and learning in the interactive classroom. Adv Physiol Educ, 2006, 30:135-140.
20. Cognitive processes in motivation [Internet]. [updated unknown; cited 2012 September 16]. Available from: http://www.education.com/reference/article/achievement-motivation/
21. Gliga M, Marusteri M. Computer Aided Concept for Enhancing Motivation and Thinking while Teaching and Learning Medical Physiology. App Med Inf, 2012, 31: 69-76
22. Xitao Fan. Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. Educ Psychol Meas, 1998; 58:357-381.