보건의료교육평가
Journal of Educational Evaluation for Health Professions
2005. 제2권 제1호, 113-118

의과대학에서 시행한 셈틀이용적응검사의 능력모수와 지필고사 성적 및 국가시험성적과 상관관계

김미영[1] 이윤환[2] 허선[1]

[1]한림대학교 의과대학
의학교육학교실 및
의학교육연구소
[2]자연대학 정보통계학과

접 수 : 2005년 4월 29일
게재승인 : 2005년 5월 17일

책임저자 : 허선
(우)200-702
강원도 춘천시 옥천동 1
한림대학교 의과대학
의학교육학교실
Tel: 033-248-2652
Fax: 033-241-1672
email: shuh@hallym.ac.kr

*이 연구는 보건복지부 보건의료 기술진흥사업의 지원에 의하여 이루어진 것임.
(03-PJ1-PG3-50300-0001)

Correlations between the scores of computerized adaptive testing, paper and pencil tests, and the Korean Medical Licensing Examination

Mee Young Kim[1], Yoon Hwan Lee[2], Sun Huh[1]
Department of Medical Education, College of Medicine and Institute of Medical Education, Hallym University, Chuncheon, Korea[1]
Department of Information and Staticstis, Hallym University, Chuncheon, Korea[2]

To evaluate the usefulness of computerized adaptive testing (CAT) in medical school, the General Examination for senior medical students was administered as a paper and pencil test (P&P) and using CAT. The General Examination is a graduate examination, which is also a preliminary examination for the Korean Medical Licensing Examination (KMLE). The correlations between the results of the CAT and P&P and KMLE were analyzed. The correlation between the CAT and P&P was 0.8013 (p=0.000); that between the CAT and P&P was 0.7861 (p=0.000); and that between the CAT and KMLE was 0.6436 (p=0.000). Six out of 12 students with an ability estimate below 0.52 failed the KMLE. The results showed that CAT could replace P&P in medical school. The ability of CAT to predict whether students would pass the KMLE was 0.5 when the criterion of the theta value was set at -0.52 that was chosen arbitrarily for the prediction of pass or failure.

Keywords: Computerized Adaptive Testing, Item Response theory, Medical Education, Evaluation, Korean Medical Licensing Examination

서 론

한림대학교 의과대학에서 2004년도 11월 제 2 차 종합시험에서 이틀에 걸친 지필고사 후 다음 날, 한 시간 동안 셈틀이용적응검사(computerized adaptive testing, 이하 적응검사)를 시행하였다. 이러한 새로운 검사법이 과연 기존의 검사법에 의한 학생 성취도 평가(또는 학습평가)와 얼마나 상관관계가 있는가 알아보려고 하였다. 적응검사는 셈틀을 이용하여 치르므로 셈틀기반검사(computer-based test)의 장점 즉, 지필고사에서 다루기 힘든 그림, 소리, 동영상을 매우 쉽게 문항에 넣어 실제 상황에 가깝게 지문이나 답가지를 제시할 수 있고 수험생의 능력을 정밀하게 측정할 수 있으며, 빠른 시간 동안 수험생의 능력을 평가할 수 있고, 수준에 맞추어 문항을 제시하여 수험생이 흥미를 잃지 않게 한다는 장점이 있다[1]. 이러한 장점과 더불어 지필고사

결과를 문항반응이론에 따라 능력모수를 산출한 결과와 적응검사의 최종 수험생의 능력모수가 일치한다면 적어도 능력을 추정하는 데는 이틀간의 지필고사를 한 시간이라는 짧은 기간의 시험으로 대치가 가능하다고 할 수 있다. 또한 이 종합시험은 졸업시험의 성격이 있을 뿐 아니라 합격 불합격을 가르는 의사국가시험의 예비시험이라는 성격이 있다. 자격시험이 준거참고검사(criteria-referenced test)로 빠른 시간 내 합격 불합격을 가리는 것이라면 적응검사가 앞으로 교육현장에 적용하는 데 효용도가 있을 것이다. 이 연구에서는 2004년도 한 의과대학 4 학년 98명 대상으로 치른 제 2 차 종합시험결과를 자료로 하여 지필고사와 적응검사에서 수험생의 능력모수값 사이의 상관관계, 지필고사의 총점과 적응검사의 능력모수의 상관관계 및 적응시험의 능력모수와 국가시험 총점의 상관관계를 알아보았다. 또한, 적응시험 결과 수험생의 능력모수가 국가시험의 합격을 예측할 수 있는 지 알아보려고 하였다.

재료 및 방법

2004년도 11월에 치른 한림대학교 의과대학 제 2 차 종합시험의 지필고사 결과 및 적응검사의 결과를 분석하였다. 수험생은 98명이었고, 지필고사는 이틀간에 걸쳐 550 문항으로 치르었다. 종료규칙은 최소 문항 수 60 문항과 능력모수 추정의 표준 오차 0.01 이하로 설정하였다. 이 최소 문항 수는 지필고사 문항의 약 1/10 수준이며, 60 문항이면 사전 검사 결과 능력모수 추정의 표준오차가 0.01 이하가 되는 것을 확인하고, 능력모수 추정에 충분한 문항수이라고 판단하였다. 종료규칙을 표준오차 0.01 이하이고 최소 60 문항을 푸는 것으로 하고, 시간 제한은 두지 않았다. 적응시험의 문항데이터베이스는 2003년도 제 1 차 및 제 2 차 임상종합시험의 문항 1,050개를 가지고 삼모수 로짓모형에 따라 ICL을 이용하여 문항난이도지수, 분별도지수 및 추측도지수를 추정하여 입력한 것을 사용하였다. 지필고사의 문항분석은 ICL(Item response theory command language)을 이용하여 하였고, 고전검사이론에 따른 총점도 구하였다[2]. 수험생의 능력모수의 범위는 -6에서 +6까지이었다. 적응검사의 수험생의 능력모수는 시험 종료 때의 능력모수로 하였다. 또한 수험생의 2005년 1 월에 치른 국가시험 총점을 보건의료인국가시험원에서 구하여 분석에 추가하였다. 종합시험 지필고사 성적이 가장 낮은 1 명이 졸업하지 못하여 국가시험은 97명의 자료를 사용하였다. 적응검사의 능력모수와 지필고사의 수험생 능력모수, 총점, 및 의사국가시험의 총점 사이의 상관관계는 dBSTAT 4.0 [3]을 이용하여 분석하였다. 의사국가시험에서 100점 만점에 평균 60점 미만을 받아서 불합격한 12명의 능력모수 및 총점을 분석하여 어떤 결과가 더 예측력이 있는 지 알아 보았다.

결 과

각 검사 결과에 대한 다중 상관분석 결과는 Table 1 과 같다. 종합시험 적응검사의 능력모수와 종합고사 지필시험의 능력모수 사이의 상관은 Fig. 1 과 같다 (r=0.8013). 종합시험의 적응검사의 능력모수와 종합고사 지필시험의 총점 사이의 상관은 Fig. 2 와 같다

Table 1.
Multiple correlation of each results of an analysis

Variable	Theta from P&P	Score from P&P	Theta from CAT	Score from KMLE
Theta from P&P	1.0000			
Score from P&P	0.9857*	1.0000		
Theta from CAT	0.7789*	0.7639*	1.0000	
Score from KMLE	0.8342*	0.8210*	0.6436*	1.0000

* : P 〈 0.05
Number of sample 97, df = 95,
significance level (α= 0.05): t = 1.9853
Abbreviation. P&P: paper and pencil test,
CAT: computerized adaptive testing,
KMLE: Korean Medical Licensing Examination.

(r=0.7861). 종합시험의 적응검사의 능력모수와 의사국가시험 지필시험의 총점 사이의 상관은 Fig. 3 과 같다(r=0.6436). 의사국가시험에서 100점 만점에서 평균 60 점 미만을 받아서 불합격한 학생의 종합시험 지필고사 총점 및 종합시험 적응검사의 능력모수를 비교한 결과는 Table 2에 있다.

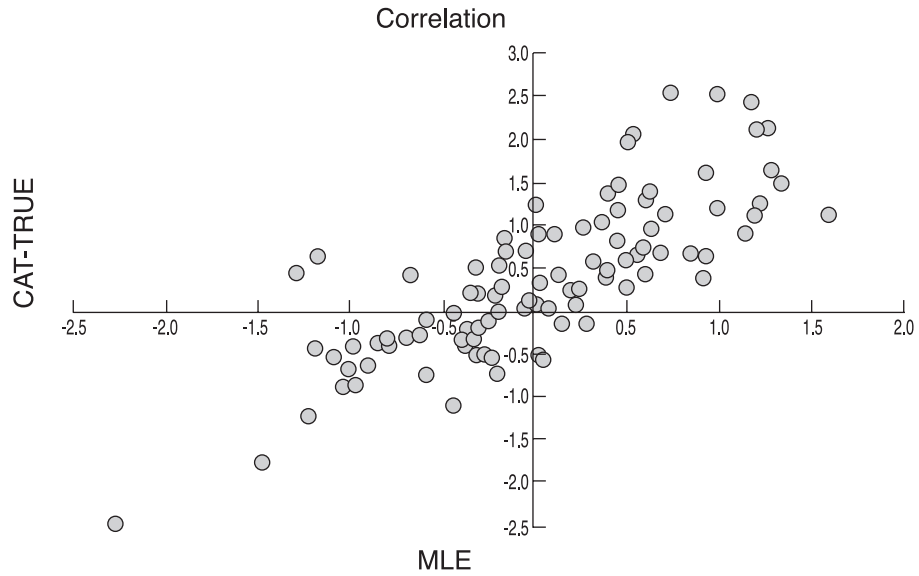

Fig. 1. Correlation between ability of examinees (CAT_TRUE) from computerized adaptive testing and ability of examinees(MLE) from paper and pencil test. r=0.8013 p<0.0000.

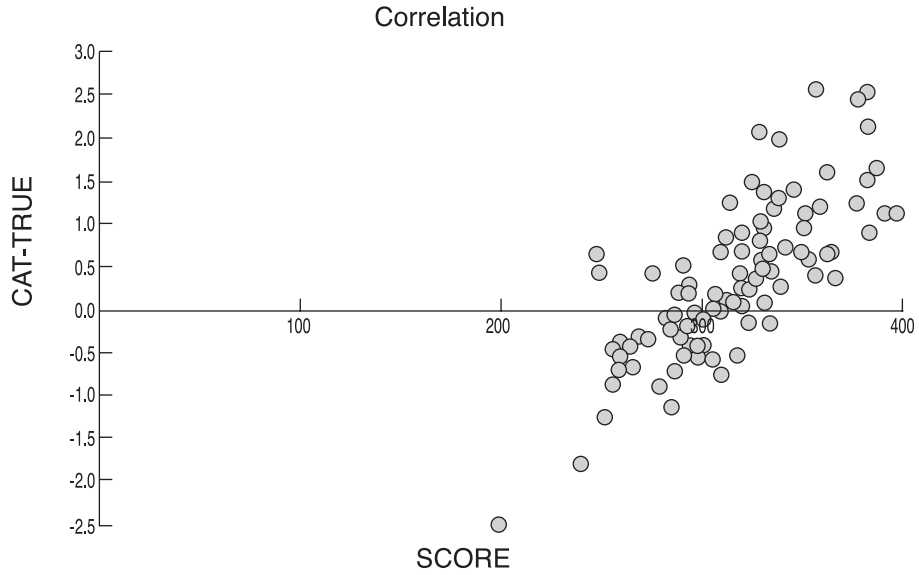

Fig. 2. Correlation between the ability of examinees (CAT-TRUE) from computerized adaptive testing and the score of examinees (SCORE) from the paper and pencil test. r=0.7861 p=0.0000.

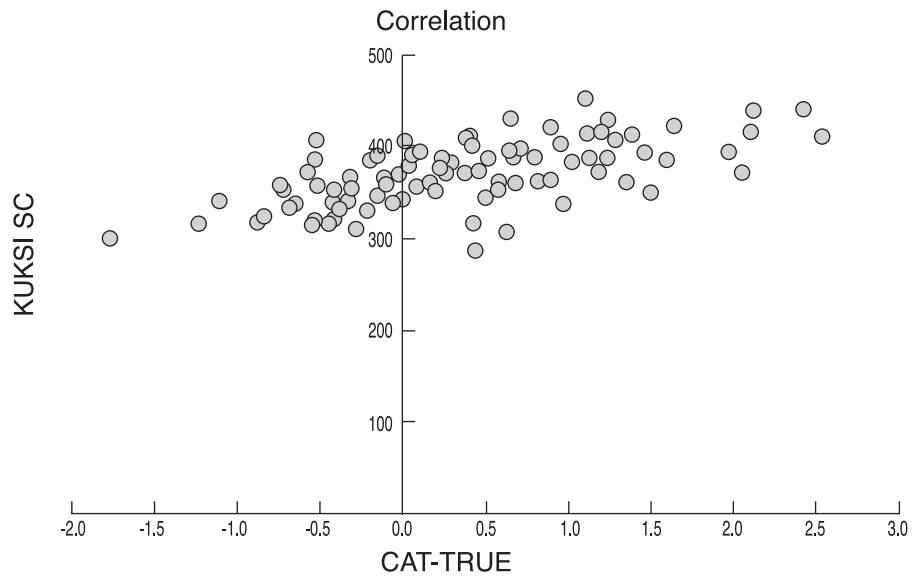

Fig. 3. Correlation between ability of examinees (CAT_TRUE) from computerized adaptive testing and score of National Medical Licensing Examination of the Republic of Korea(KUKSI SC) from paper and pencil test. r=0.6436, p<0.0000.

Table 2. Comparison of the score of paper and pencil test(P&P), ability estimate(theta) of computerized adaptive testing(CAT) from failed students in Korean Medical Licensing Examination(KMLE)

P&P score(550 items)	CAT theta
248	0.628
249	0.437
276	0.422
273	-0.281
264	-0.409
257	-0.442
298	-0.531
260	-0.542
256	-0.843
279	-0.874
251	-1.233
240	-1.771

고 찰

위와 같은 결과는 이번 제2 차 종합시험의 적응검사에 따른 능력모수는 지필고사의 능력모수와는 높은 상관이 있으나(r=0.8013, p=0.000), 지필고사의 총점과의 상관(r=0.7861, p=0.000) 및 국가시험의 총점과의 상관은 그보다는 낮음을 알 수 있다(r=0.6436, p=0.000). 제 2 차 종합시험의 지필고사 뒤 바로 치른 적응시험의 능력모수와의 상관이 지필고사 총점의 상관이나 그보다 2 달 열흘 이후에 치른 의사국가시험의 총점과의 상관보다 높은 것은 당연한 결과이다. 즉 추정 방법이 같은 도구를 쓸 때, 총점보다는 능력모수와의 상관이 높아야 할 것이다.

그렇다면 이런 적응시험의 효용도는 무엇일까? 즉, 가장 바람직하기는 적응시험으로 의사국가시험의 합격 불합격 여부까지 유추가 가능한 것이다. 그러나 이번 결과에서는 능력모수가 -0.52 미만인 최하위 집단에서 12 명 중 6 명이 불합격하고 6 명이 합격하였다. 물론 능력

모수가 낮은 사람이 총점도 낮지만 국시 합격여부를 적응시험으로 예측하는 것은 최하위 집단을 제외하면 쉽지 않다. 능력모수 최하위 5 명 중 4 명이 불합격하였다. 지필고사의 총점이 하위 12명의 상한점수인 266점을 기준으로 하면 12명 중 8 명이 불합격하였다. 이런 면에서 보면 지필고사의 총점이 불합격을 예측하는 데 더 나을 수 있으나, 여기서 비교한 의사국가시험의 성적은 고전검사이론에 의한 총점이다. 만약 의사국가시험을 문항반응이론으로 분석하여 각 개인의 능력모수와 종합시험의 능력모수를 비교한다면 더 높은 상관과 더 높은 예측력을 가질 수 있을 것이다.

적응시험을 학교 단위나 과 단위에서 적용하는 것은 예가 많지 않다. 적응시험의 적용 대상은 일년 내내 결과가 필요한 경우로 국가단위 면허시험, 교실 단위 진단시험, 미육군선발시험과 같은 것이 있고, 셈틀을 이용하여 시험 보는 것이 내용에 가장 적당한 경우 예로 CAD 자격시험 같은 것이 있다. 한편 학기제 수업 평가, 일년에 한번 시험보는 경우, 수행평가 등에는 적당치 않다고 한다[4, 5]. 그렇다면 의사면허 시험과 같은 자격시험에는 적당할 수 있으나 그 자격시험에 대한 예비시험 성격인 의대 4학년 대상 임상종합시험은 지속하여 몇 차례 시험을 보는 것이므로 문항 데이터베이스만 충분하면 좋은 대상이 될 수 있다.

지필고사의 능력모수와 적응시험의 능력모수의 상관이 0.8013이라는 것은 매우 높은 상관이다. 이 정도면 적응시험의 진점수로 지필고사의 총점을 대신하여도 큰 문제는 없는 수준이다. 그렇지만 더 높은 상관을 위하여 이번 종합시험의 적응시험의 몇 가지 한계를 지적하고 앞으로 개선이 필요함을 제안한다. 첫째, 이번 시험의 문항데이터베이스가 2003년도 종합시험 문제이므로 이미 다 노출된 문항이었다. 이렇게 노출된 문항으로도 능력의 변별은 가능하였고, 노출된 문항을 수험생이 모두 다 익히고 있지는 않으므로 전혀 문항의 수정이 없어도 모두 맞추는 것이 아니었기 때문에 노출된 문항도 능력모수의 추정에는 문제가 없다고 할 수 있으나 앞으로 문항 데이터베이스의 수를 충분히 늘려 노출에 의한 능력모수 추정의 삐뚤림(bias)을 최소로 하여야 한다. 두번째는 삼모수로짓모형을 도입하여 적응검사를 치르었으나, 삼모수로짓모형을 도입하기 위하여 최소 500 명 정도의 수험생을 통하여 문항모수를 추정하는 것이 필요하다고 한다. 한림대학교 의과대학은 한 학년이 80 명 수준이므로 이 문항모수 추정의 정밀도에 문제가 있을 수 있다. 그러나 그대로 시행한 이유는 일차원성 검증이 가능하였고, 적합도 검사(goodness of fit) 에서 부적합 문항이나 부적합 수험생 수가 적었기 때문이다(자료 미제시). 이렇게 수험생 수가 적을 때는 다른 학교와 연대하여 시험 본 후 문항모수를 추정하는 것도 한 방법이고 또는 라쉬모형을 사용하여 수험생수가 적어도 문항모수 추정에 문제가 없도록 하는 방법도 있다.

이런 추정에 대한 내용 이외에도 내용 면에서 과거의 시험은 모두 지필고사라 멀티미디어 문항이 적어서 셈틀을 이용한 시험의 장점을 살리지 못하였다는 점이 있다. 이런 여러 문제는 앞으로 문항 데이터베이스를 10,000 개 이상으로 수를 늘려서 노출 효과를 최소로 하고, 라쉬모형의 도입을 검토 또는 시행하여 삼모수로짓모형과 비교하고, 문항 개발할 때 멀티미디어 문항 개발을 장려하여 멀티미디어 자료를 시험에서 활용하게 하는 등으로 극복하여야 할 것이다.

이렇게 우리 나라에서 대학의 학과, 학년 단위의 시험에서 적응검사를 시행하는 예는 부산대학교 교육학과를 제외하고 매우 드물다[6]. 특히 의대의 학부 교육과정에서는 전례를 찾아 볼 수 없다. 앞으로 적응검사를 의학교육평가에 도입하는 것이 학생 교육의 질 향상이나 시험 문항이나 평가의 질 향상에 얼마나 기여할 수있을지 연구가 필요하다.

결 론

적응시험은 지필고사를 대체할 수 있을 만큼 적응검사의 능력모수와 지필고사의 총점 사이에 높은 상관있었다. 적응검사는 의사국가시험과의 상관이 지필고사보다 떨어지고 불합격에 대한 예측도도 떨어지지만, 앞으로 의사국가고사를 문항반응이론으로 분석한 결과와 비교하면 더 나아질 수 있을 것이다. 이런 적응검사의 문항특성값의 추정을 안정되게 하기 위하여 앞으로 문항 데이터베이스 구축할 때 분석 대상 수험생 수를 충분히

확보하거나 라쉬모형을 도입하는 해결책이 가능할 것이다.

참고문헌

1. 부재율. 컴퓨터활용검사. 교육과학사: 서울 2002.
2. ICL, Item response theory command language [computer program] available from http://ssm.sourceforge.net
3. Kim SN, Huh S. Can Statistics used in the Medical Journals from Korea be Covered by Korean Statistical Program, dBSTAT? Korean J Med Educ. 2002; 14: 111-117.
4. Van der Linden WJ, Glas CAW. Computerized adaptive testing Theory and Practice. London, U.K: Kluwer Academic Publishers, 2003.
5. Wainer H. Computerized adaptive testing: A Primer. 2nd ed. Mahwah, New Jersey, U.S.A: Lawrence Erlbaum Assoc. 2000.
6. 김영환, 손미, 정희태. 컴퓨터기반 적응검사(CAT)의 이론과 실제. 서울: 문음사, 2002.