

# 의과대학 한 과목 시험에서 문항반응이론에 따른 검사 동등화 예

한림대학교 기생충학교실 및 의학교육연구소

허 선

= Abstract =

## Test Equating of a Medical School Lecture Examination Based on Item Response Theory: A Case Study

Sun Huh

*Department of Parasitology, College of Medicine and Institute of Medical Education, Hallym University, Chuncheon Korea*

**Purpose:** To determine possible differences in the ability of students from two consecutive medical school class years, test equating based on item response theory was performed on the results of a lecture examination.

**Methods:** The dataset for this study of the results of a medical school lecture exam was composed of the number of items and examinees: 75 students tested on 60 items in 2002, and 82 students tested on 50 items in 2003. We used common item non-equivalent group design and tested the assumption of unidimensionality and data-fitness. Item parameter estimates were based on the Rasch model, and the scale transformation was performed using mean/sigma moment methods. Equating was applied to both true and observed scores, and the results were analyzed.

**Results:** The slope and intercept for the scale transformation of the 2003 exam to the 2002 exam were 1.2796 and 0.9630, respectively. The mean and standard deviation of the ability parameter of the students taking the 2003 exam were 1.69 and 0.66, respectively. After scale transformation, the mean increased to 3.13 and the standard deviation increased to 0.85. For the 2002 exam, the mean was 0.70 and the standard deviation was 0.66. The correlation coefficient between the true and observed scores was 0.9994 ( $P = 0.00$ ).

**Conclusion:** The difference in the ability parameter of students between the two years increased after scale transformation. Both the true and observed scores could be used interchangeably after equating. Thus, test equating based on item response theory may be applicable to medical school lecture examinations.

---

**Key Words:** Medical education, Evaluation, Test equating, Item response theory, Rasch model

---

교신저자: 허 선, 한림대학교 의학교육연구소  
강원도 춘천시 옥천동 1

Tel: 033)248-2652, Fax: 033)251-8354, E-mail: shuh@hallym.ac.kr

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

## 서 론

왜 시험을 치른 뒤 검사 동등화 (test equating)가 필요할까? 문항반응이론에 따르면 어느 문항이나 고유의 특성이 있어 수험생 (examinee) 집단이 다르더라도 특성이 바뀌지 않고, 마찬가지로 수험생도 고유의 능력이 있어 문항 특성에 따라서 능력이 바뀌지 않는다. 그렇지만, 현실은 문항모수 (item parameter) 추정 시 각각 다른 집단에서 시행한 것이라서 그 집단에서 능력모수 (ability parameter) 또는 난이도모수 (difficulty parameter)를 임의로 평균 0, 표준편차 1을 지정하고 추정하기 때문에 만약 어느 시험을 치른 두 집단을 일정한 모집단에서 추출한 집단일 경우라도 두 집단의 능력이 차이가 나면 문항 특성이 각각 다를 수 있으므로 동등화가 필요하다 (남현우, 2001). 문항모수는 문항특성곡선 (item characteristic curve)에서 찾을 수 있는 난이도모수 (difficulty parameter), 분별도모수 (discriminating parameter), 추측도모수 (guessing parameter)와 같이 문항의 특성을 나타내는 모수이다. 난이도모수는 문항의 어려운 정도를 나타내는 지수로 문항반응이론의 1모수 및 2모수 모형에서 난이도모수는 어떤 문항을 맞출 확률이 0.5에 해당하는 수험생의 능력수준의 점을 의미한다. 값이 클수록 어려운 문항이라고 해석한다. 분별도모수는 어느 문항이 능력이 높은 사람과 낮은 사람을 잘 구별하여 낼 수 있는지를 파악하는 모수로 문항특성곡선에서 문항난이도에 해당하는 점에서의 기울기로 나타난다. 값이 클수록 분별력이 높다고 해석한다. 삼모수 모형에서 쓰이는 추측도모수는 선택형 문항에서 수험생이 추측에 의하여 문항의 답을 맞출 확률을 뜻하는 모수이다. 오지선다형 문항에서 추측도모수는 0.2 전후로 추정하나 문항마다 답가지의 내용에 따라 다르다. 능력모수는 문항반응이론에 의하여 검사에서 측정된 수험생의 잠재 능력에 대한 모수를 추정된 값으로  $\theta$ 로 표기한다. 이 값은 대개 -3에서 +3까지 또는 -6에서 +6까지로 표기할 수 있고, 높은 값일수록 수험생의 능력이 높음을 뜻한다. 일모수 모형이나 라쉬모형 (Rasch)모형에서는 문항모수 가운데 분별도는 각 문

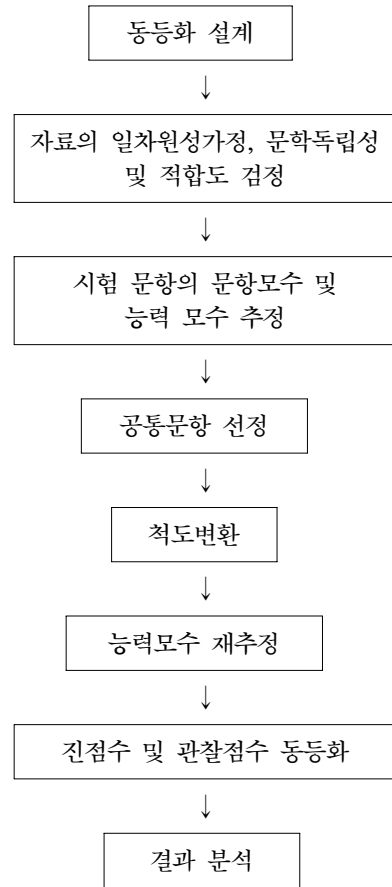


Fig. 1. Flow chart for the test equating process

항이 같으며, 추측도모수는 없다고 보고 난이도모수만 추정을 한다.

의과대학에서 시행하는 각 시험 결과의 해석에는 몇 가지 고려하여야 할 필요가 있다. 우선 매년 평균이 다른데 일정한 수준이 아니고 편차가 심할 경우가 있다. 그러면, 그 원인은 수험생의 능력이 갑자기 좋아지거나 나빠졌는지 아니면 문항의 난이도 조절에 실패하였는지 점검하여야 해야 한다. 왜냐하면 의과대학의 시험은 단순히 학점을 매기기 위한 목적 이외에 최소 기준에 미달한 수험생은 낙제를 시키는 준거참조평가 (criterion-based test)의 역할이 있기 때문이다. 수험생의 학습 능력의 차이가 아니라 문항의 난이도 차이로 매년 낙제시키는 수험생의 수가 차이가 난다면 공평하지 못하다는 불만

Table I. General Characters of two Examinations in 2002 and 2003

Year	2002	2003
Number of items	60	50
Number of examinees	75	82
Mean	37.95 (63.3%)*	38.65 (77.3%)
Standard deviation	5.49	3.30

\* % value is the ratio of score to score 100.

을 들 수 있다. 그러므로, 한 의과대학의 한 교실에서 시행한 2002, 2003년도 필기 시험 결과의 차이가 과연 각 연도별 수험생의 능력의 차이인지 아니면 단순히 난이도 조절의 문제였는지 알아보기 위하여 문항반응이론에 따라 2003년도 문항을 2002년도 문항에 동등화작업을 시도하였다. 고전검사이론에 따른 동등화 방법도 있는 데 문항반응이론을 도입하는 이유는 우선 일차원성 가정 및 자료 적합성을 성립하면 여러 모수 측정을 정밀하게 할 수 있다는 점 뿐 아니라, 이런 작업 결과를 샘플이용 적응 검사(computerized adaptive test)에 사용하는 문항 데이터베이스에 입력을 전제로 하기 위한 때문이다. 문항 데이터베이스를 구축할 때 문항 특성을 입력한다면 각 연도별로 추정된 문항모수치를 입력하는 것은 매년 문항의 난이도 차이로 인하여 각 특성 값에 표준화가 어렵다. 그럴 경우에도 역시 동등화 작업을 시행하여 연도별 차이를 보정하여 주어야 한다. 이렇게 의과대학 개설 교과목 단위 과목의 시험에 대하여 문항반응이론에 따른 동등화 작업을 시도한 예는 코리아메드 (<http://koreamed.org>), 퍼브메드 (<http://pubmed.org>) 및 Web of Science (<http://isi02.isiknowledge.com/>) 를 검색하여 본 결과 국내외 어디서도 찾을 수 없다. 그 이유는 문항반응이론을 도입하기 위한 가정 (assumption)인 일차원성 가정 (assumption of unidimensionality) 및 자료 적합성(data-fit)에 대한 검정을 하기 어려워질 것으로 추측한다. 또한 그 쓰임새를 의학교육 평가에서 구체로 지적하기 어려워질 것이다.

문항반응이론에 따른 검사 동등화 과정은 크게 세 가지 절차를 거치는 데 첫째는 문항모수의 각각

추정, 둘째, 두 시험 문항모수의 척도 변환(scale transformation)이며, 셋째는 진점수(true score) 또는 관찰점수(observed score) 동등화 과정이다 (Kolen and Brennan, 1995). 문항반응이론에 따른 동등화 방법 중 문항반응이론에 따른 방법을 사용할 때 라쉬모형(Rasch model)을 따르거나 삼모수 로짓모형(three parametric logistic model)을 따르는 방법이 있다. 이번 증례에서는 문항수가 50, 60개이며 수험생수가 100 명이 넘지 않으므로 라쉬모형을 적용하였다(지은림 및 채선희, 2000). 만약 샘플이용 적응검사에서 필요한 문항모수의 동등화를 원한다면 두 번째인 척도 변환 과정으로 충분하다. 척도변환은 검사 동등화와 혼용하여 쓰이기도 하므로, 이 과정만으로도 문항 모수의 동등화라고 할 수 있다. 또한 성적을 능력모수만으로 나타낸다면 더 이상의 동등화 과정은 필요 없다. 척도 변환 후에 시행하는 점수 동등화 과정은 능력모수만으로 원점수와 비교가 금방 되지 않으므로 능력모수를 원점수(raw data)와 비교 가능한 자료로 만들어 준다. 점수동등화에서는 진점수 및 관찰점수 동등화가 있다. 진점수 동등화 과정은 검사특성곡선을 이용하여 두 시험의 맞은 점수와 관계를 보여 주며, 어떤 주어진 능력모수 값이 두 시험의 진점수와 동등하게 된다. 관찰점수 동등화는 관찰한 맞은 점수의 분포를 추정하고 그것을 동백분위 방법으로 동등화하는 것이다(남현우, 2001). 이런 동등화 과정을 Fig. 1과 같이 모식도로 그릴 수 있다.

그러므로, 우선 일개 시험 과목에서도 문항반응이론을 도입할 수 있는 일차원성 가정 및 자료 적합성을 검정할 수 있는 지 알아보고, 그 후 동등화 작업을

을 알려진 방법의 절차를 따라 시행하여 단위 과목에서 동등화 작업의 효용성을 검토하였다. 이런 결과는 앞으로 일개 과목 시험에서 뿐 아니라, 국가시험과 같은 대규모 시험 (high stake examination)의 동등화 작업에 적용하는 데 참고 자료로 삼을 수 있을 것이다. 이런 작업 결과를 가지고 수험생의 연도별 능력의 차이가 성적의 차이를 가져 온 것인지 점검할 수 있을 뿐 아니라, 샘플이용 적응검사에서 문항 데이터베이스에 문항모수 입력에도 활용할 수 있다. 또한 이러한 작업은 문항반응이론을 의학교육평가 분야에 도입하는 한 예가 될 것이다.

## 대상 및 방법

### 1. 대상

한 의과대학의 기초 의학 과목 중 하나로 2002년도 및 2003년도 시험 결과를 대상으로 하였다. 대상 시험 문항 수는 2002년도 60, 2003년도 50문항이었다. 수험생 수는 2002년도 75명 2003년도 82명이었다. 시험은 매년 누리집보기 폴그림 (web browser)를 통하여 시험 문항에 접근하여 수험생이 응답하고 마지막에 결과를 서버 (server)로 보내는 샘플기반시험 (computer based test)으로 치르었다. 시험 도중 수험생은 자신의 응답 결과를 자유롭게 수정할 수 있도록 하였다. 결과는 맞으면 1, 틀리면 0으로 표시하는 이분문항으로 변환하여 각 응답자 및 문항별 행렬 자료로 확보하여 다음 작업에 사용하였다 (부록 1).

2003년도 시험을 2002년도에 동등화하려고 하였다. 2002년도 필기시험은 문항이 총 100문항이었지만 40 문항은 단순히 학명을 물어 보는 문항으로 대부분의 수험생이 다 맞추고, 또한 2003년도에는 빠진 내용이므로 분석에서 제외하였다. 2002년도 및 2003년도의 문항수, 수험생수, 평균 및 표준편차는 Table 1과 같다. 앞으로 2002년도 시험을 X, 2003년도 시험을 Y로 놓는다.

### 2. 동등화 설계

동등화의 여러 설계 중 공통문항 비동등집단 설

계 (common item non-equivalent group design)를 사용하고, 2002년도 및 2003년도의 수험생의 집단 같은 모집단에서 선택한 표본이 아닌 독립된 표본으로 보고, 공통 문항을 이용하여 동등화를 시도하였다.

### 3. 일차원성, 문항독립성 및 적합도 검증

문항반응이론에 따라 문항모수 및 능력모수 (ability parameter)를 추정하기 위하여는 검사의 일차원성 (unidimensionality) 및 국소 독립성 (local independence)가정을 검정하여야 한다. 또한 문항모수 및 능력모수의 적합도 (goodnes of fitness)를 살펴보아 심하게 어긋나는 문항이나 수험생이 있는 지를 살펴 보아야 한다. 일차원성검정은 어느 시험이 단일한 차원을 물어 보는 지 예를 들면 해부학 시험이면 해부학에 대한 능력을 측정하는 것이지, 해부학 시험을 치르기 위하여 영어 실력이나 한자 실력이 있어야 지문을 해독할 수 있다면, 그것은 다른 차원의 능력을 측정한다는 것이다. 또한 국소독립성은 한 문항이 다른 문항에 답하는 데 영향을 미치지 않고 각 문항이 모두 독립적이어야 한다는 것이다. 이 일차원성 가정이나 국소독립성 가정을 만족시키기 위하여는 우선 문항을 살펴서 모든 문항이 측정 목적에 맞는지를 점검하고 또한 어느 문항이 다른 문항의 해결에 영향을 주는 암시나 자료 제시가 있는 지를 점검하여야 한다. 그 외에 요인분석 (factor analysis) 등을 통하여 한 시험이 몇 개의 요인으로 구성되어 있는 지 점검하는 통계 분석을 시도한다. 그러므로 이번 분석에서는 면밀히 문항 내용을 살펴 목적에 맞는 지와 다른 문항에 영향을 줄 문항이 있는 지를 관찰하고 Holland Rosenbaum's Method (=HR-MH) (Nandakumar, 1993)를 이용하여 일차원성 가정을 검정하였다. 문항독립성 가정은 각 문항이 다른 문항의 해결에 영향을 미치지 않은 내용이 독립된 질문이므로 성립한다고 간주하였다. 적합도 검증은 라쉬모형에 따른 문항모수 추정 프로그램 (program)인 BIGSTEPS를 사용하여 적합하지 않은 문항이나 수험생이 얼마나 있는 지 찾아보았다.

**4. 연도별 시험의 문항모수 및 능력 모수 추정**

라쉬모형을 기반으로 각 문항의 난이도모수에 따라 능력모수를 추정하는 BIGSTEPS를 사용하였다 (부록 2).

**5. 공통문항 선정**

2002년도 시험 및 2003년도 시험 문항을 비교하여 그 유형 (type) 및 수준 (level)이 같고, 내용이 같은 문항을 선정하여 라쉬모형에 따른 난이도의 상관관계를 분석하였다. 분석 후 멀리 떨어진 자료 (outlier)를 제외하고 다시 상관관계를 구하였다. 그렇게 구한 각 연도의 문항을 공통문항으로 선정하여 척도 변환에 사용하였다.

**6. 척도 변환**

두 시험 X (2002), Y (2003) 의 내용을 점검하여 그 유형 (type) 및 수준 (level)이 같고, 같은 주제를 물어 본 문항을 공통 문항으로 선정하였다. 그러므로, 이 경우 공통문항이라기보다는 가상공통문항 (virtual common item)이라는 표현이 더 적당하다.

공통 문항으로 우선 선정한 22 문항의 라쉬모형에 따라 구한 문항난이도를 짝 비교하여 중심에 해당하는 직선 그림을 그려본 후, 직선에서 매우 멀리 떨어진 문항을 제외하고 14 문항의 짝을 가지고 다시 그림을 그려 기울기와  $\chi$ 절편을 구하였다. 이 경우 각 짝 사이의 값 차이가 적어야 하는 데, 그 수준을 라쉬모형에 따른 문항분석 결과 임의로 1이하인 문항만 선정하였다. 그리고, Y의 문항모수를 X의 문항모수를 참조하여 ST를 이용하여 척도변환하고 ST에서 나오는 4가지 해구하는 방법 중 중에서 평균/표준편차에 의한 방법의 결과를 사용하였다. ST를 이용하는 것은 3모수로짓모형에서 가능하나, 모든 문항의 분별도모수를 1, 추측도모수를 0으로 놓고 계산하면 라쉬모형의 척도변환에 사용할 수 있다. 이렇게 하면 Y시험은 X시험의 참조틀 내에 있다. ST 통제 파일은 부록 3에 기술하였다.

**7. 척도 변환 뒤에 2003년도 수험생의 능력모수 재추정**

척도 변환을 하여 구한 기울기 및 절편을 이용하여 2003년도 문항모수의 척도를 주고 능력모수를 BIGSTEPS를 이용하여 재추정하였다. Y를 X에 문항 척도 변환한 결과를 가지고 재 분석 시, BIGSTEPS 플그림에서

$$USCALE = ST에서\ 구한\ 값의\ slope$$

$$UMEAN = ST에서\ 구한\ 값의\ intercept$$

이렇게 설정하고 재분석을 한다. USCALE (로짓의 척도 값) 및 UMEAN (문항척도의 중심값)을 추가하고 다시 분석한 Y시험에서 수험생의 능력 모수 및 문항 모수를 구한다. 재추정결과를 추정 전의 능력 모수 및 2002년도 수험생의 능력모수와 비교하였다.

**8. 진점수 및 관찰점수 동등화**

진점수 및 관찰점수 동등화는 능력모수를 수험생의 성취도로 표시할 때는 구할 필요가 없다. 척도 변환하여 다시 능력모수를 추정하는 것으로 동등화 과정은 마친다. 그러나 이렇게 진점수 및 관찰 점수 동등화를 시도하는 이유는 일반인이 이해하기에 점수로 표시하여야 이해하기 편하고, 또한 수험생도 마찬가지이기 때문이다. 이 동등화는 PIE를 사용하여 구하였다. PIE의 통제 파일은 부록 4와 같다.

**9. 결과 분석**

결과를 각 방법에 따라 DBSTAT (김수영 및 허선, 2002)를 이용하여 분석하였다.

**결 과**

**1. 일차원성 가정 및 문항독립성 가정**

시험의 내용을 파악하여 해당 과목과 무관한 내용은 없는 것을 확인하였다. 또한 다른 문항의 정답 선택에 영향을 줄 수 있는 문항도 없는 것을 확인하였다. 공통 문항의 선정에서 문항이 완벽하게 같은 것은 출제하지 않고, 같은 내용을 다른 지문으로 물

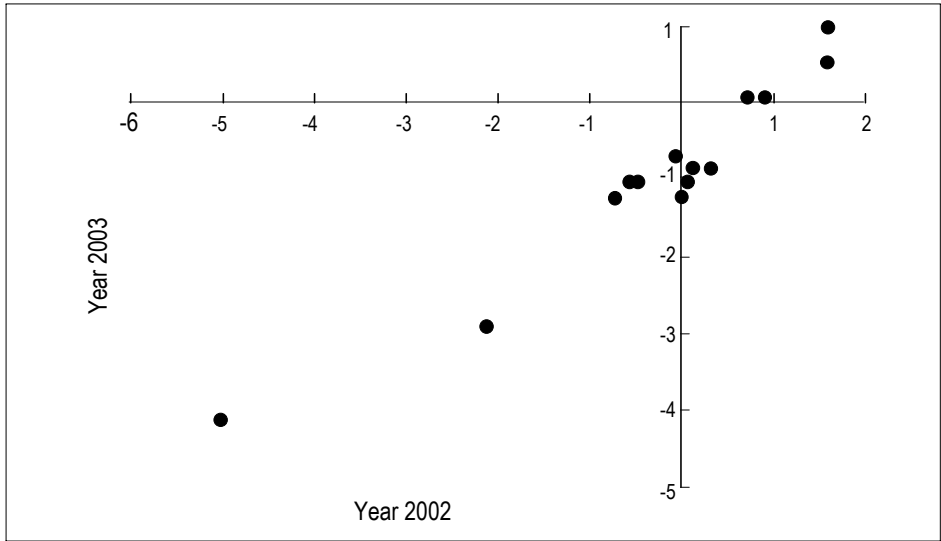


Fig. 2. Correlation between the difficulty index of common items from two examinations in 2002 and 2003

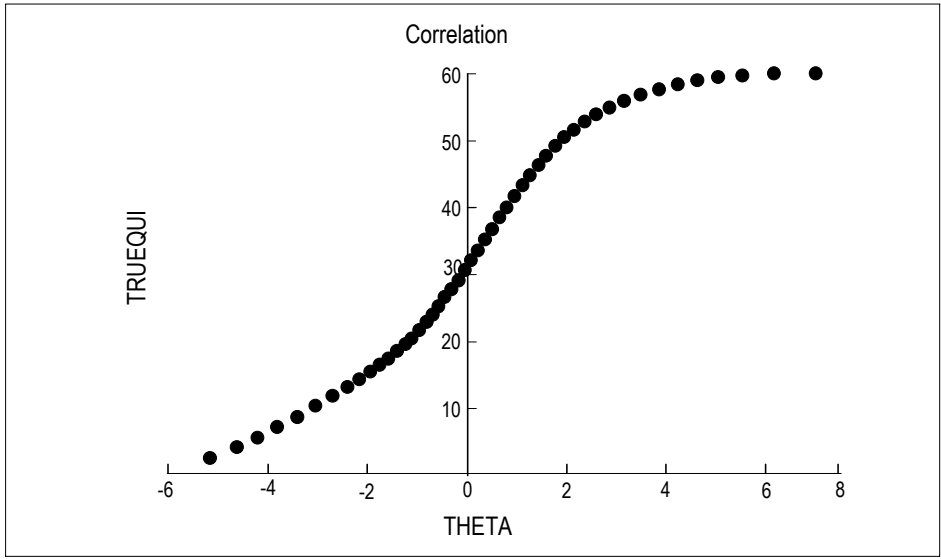


Fig. 3. Correlation between theta(ability parameter) of 2003 exam and true score equivalent to 2002 exam

어 보거나 답가지를 달리 하거나 답가지 번호 순서를 달리 한 것이라도 문항의 형태 및 수준이 같고 측정하려는 내용이 같으면 공통문항으로 하여 우선 선정하여서 상관관계 분석하였다.

일차원성가정을 검정하기 위한 HRMH의 수행 결과 2002년도 및 2003년도 시험은 다음과 같았다.

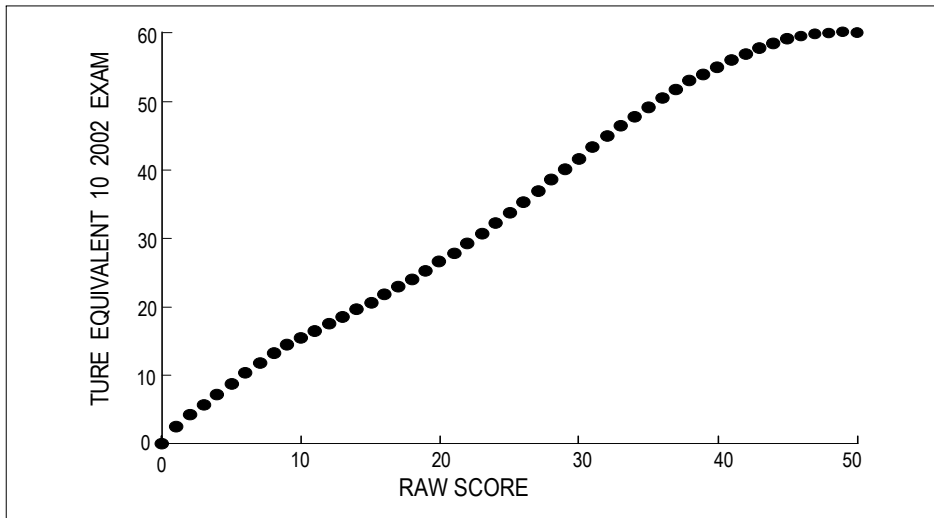


Fig. 4. Correlation between true score equivalent to 2002 exam and raw score of 2003 exam

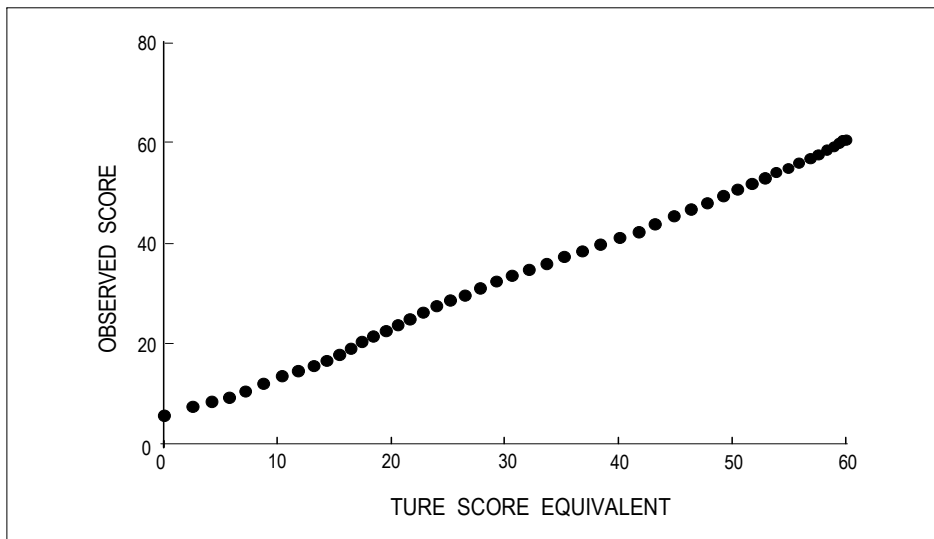


Fig. 5. Correlation between true score and observed score of the examination in 2003

2002년도

PROPORTION OF SIGNIFICANT ZS AT LEVEL  
.01 = 1.0/ 1770

PROPORTION OF SIGNIFICANT ZS AT  
LEVEL .05 = 16.0/ 1770

PROPORTION OF SIGNIFICANT ZS AT  
LEVEL .10 = 46.0/ 1770

2003년도

PROPORTION OF SIGNIFICANT ZS AT LEVEL  
.01 = .0/ 1225

PROPORTION OF SIGNIFICANT ZS AT  
LEVEL .05 = 12.0/ 1225

PROPORTION OF SIGNIFICANT ZS AT  
LEVEL .10 = 22.0/ 1225

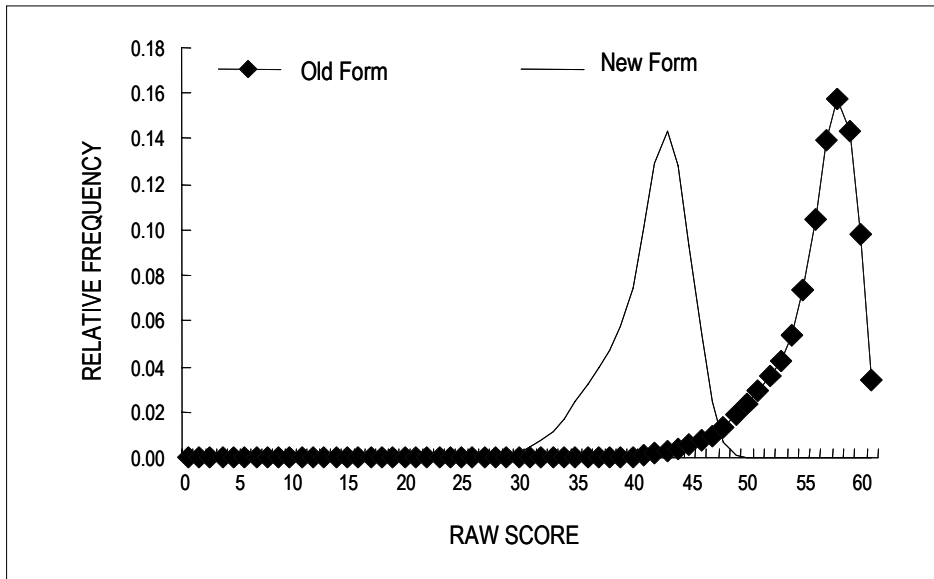


Fig. 6. Relative frequency distribution of the observed score of the examinations in 2002 (old form) and 2003 (new form) to raw score

## 2. 자료의 적합성

BIGSTEPS를 이용하여 라쉬 모형에 따른 문항모수 및 능력모수를 각각 구한다 (지은림과 채선희, 2000). 라쉬모형에 대한 적합성을 검정하기 위하여 BIGSTEPS로 분석한 결과 비적합 문항 수가 2002년도에는 60문항 중 3문항이 ZSTD (mean square fit statistics를 평균 0, 변량 1로 표준화한 값)값이 2.0을 넘었다. 2003년도 시험 결과를 BIGSTEPS로 분석하여 본 결과 적합도 통계치 (ZSTD) 50문항 중 4문항이 -2.0이하이거나 +2.0이상이었다.

## 3. 공통 문항 선정

완벽하게 같은 문항은 없으므로 비록 지문은 틀리지만 물어 보는 내용이 같은 것을 공통 문항으로 선별하였다. 이 공통문항은 양 시험 문항에서 공통으로 출제된 문항이라는 개념이다. 선별 결과 22문항을 골라 낼 수 있었다. 두 시험의 공통문항의 난이도 지수를 그림으로 그려보아서 중심선을 그린 뒤, 중심선에서 멀리 떨어진 것을 제외하고 다시 14문항으로 그림을 그렸다. 기준은 두 난이도 지수의

차이가 1.0미만인 것만 선정하였다. 14문항으로 그린 그림은 Fig. 2와 같다. 상관관계를 구한 결과 상관계수는 0.9678 ( $P = 0.0000$ )으로 양의 상관관계가 있었다. 2002년도의 공통문항의 난이도 평균 -0.2571, 표준편차 1.6739이며, 2003년도 공통문항의 난이도 평균 -0.9536, 표준편차 1.308이었다.

## 4. 공통문항을 이용한 척도 변환 및 2003년도 수험생의 능력모수 재추정

ST를 이용한 척도 변환의 결과 중에서 평균/표준편차법에 따르면 기울기 1.2796, 절편 0.963으로 2003년도 문항이 2002년도 문항과 공통 척도로 놓인다.

여기서 척도 변환 후 2003년도 시험 문항모수 및 능력모수의 재추정을 위한 control 파일에서

USCALE = 1.2796 ; 문항 난이도 로짓의 척도값 설정. 기본값은 1

UMEAN = 0.9630 ; 문항 난이도 척도의 중앙값을 설정. 기본값은 0

이와 같이 명령어를 추가하고 추정하였다.



### 5. 척도변환 전후 능력모수 비교

척도변환 전의 2003년도 수험생의 능력모수의 평균은 1.69 표준편차 0.66이었는데, 척도 변환 이후의 2003년도 수험생의 능력모수 평균은 3.13, 표준편차 0.85로 증가하였다. 2002년도 수험생의 능력모수의 평균은 0.70 표준편차 0.66 임에 비추어 척도 변환 전이나 후 모든 값에서 2003년도 수험생의 능력이 2002 년도 수험생의 능력보다 높음을 알 수 있다. 척도변환 후의 2003년도 수험생의 능력모수 표준오차는 0.095, 변환 전은 표준오차 0.073, 2002년도 수험생의 표준오차는 0.076 이었다.

### 6. 진점수 및 관찰점수 동등화

PIE를 이용한 동등화한 결과 2003년도 진점수와 능력모수 사이의 상관관계는  $r = 0.9668$  ( $p = 0.0000$ , Fig. 3), 2003년도 진점수와 원점수 사이의 상관관계는  $r = 0.9959$  ( $p = 0.0000$ , Fig. 4), 2003년도 진점수와 관찰점수 사이의 상관관계는  $r = 0.9994$  ( $p = 0.0000$ , Fig. 5), 2003년도 능력모수와 원점수 사이의 상관관계는  $r = 0.9810$  ( $p = 0.0000$ ) 이었다. 2003년도 및 2002년도의 각 관찰점수의 원점수에 대한 상대빈도는 Fig. 6 과 같다.

## 고 찰

의과대학의 일개 시험 과목에서도 문항반응이론을 도입하기 위한 전제인 일차원성 가정 및 자료 적합성을 점검하여 본 결과, 충분히 라쉬모형에서 일차원성 가정 및 자료 적합성을 충족시킴을 알 수 있었다. 일차원성 가정에서는 HRMH를 이용하였을 경우 각 문항사이의 상관관계가 유의수준을 벗어나는 조합이 충분히 적었다. HRMH에 따르면 모든 유의 수준에서 일차원성 가정을 만족한다. 어느 유의수준에서 유의한 Z 값의 비율(PROPORTION OF SIGNIFICANT ZS AT LEVEL)의 값이 그 유의수준보다 적은 값이면 일차원성 일차원성 가정을 만족한다고 본다. 예로 유의수준 0.01일 때 값이 0.01 이하일 때, 유의수준이 0.05 이면 이 값이 0.05 이하

일 때 유의수준이 0.1 이면 이 결과 값이 0.05 이하일 때 일차원성 가정을 만족시키는 것으로 해석한다. 또한 이러한 검정 결과는 문항 내용이 주어진 과목에 해당하는 문항으로 구성되어 있지, 다른 과목의 내용으로 유추할 수 있는 것은 없었다는 점도 일차원성 가정을 뒷받침하는 내용이다. 각 문항 내용의 독립성은 전문가 판단에 따라서 어떤 문항이 다른 문항의 답을 암시하거나 영향을 미치지 않음을 판단하는 것으로 검정하였다.

자료 적합성 중 우선 문항 적합성은 부적합 문항이 얼마나 되는 지 파악하여야 하는 데 2002년도에는 60문항 중 3문항이 ZSTD (mean square fit statistics를 평균 0, 변량 1로 표준화한 값)값이 2.0을 넘어 소수이므로 대부분 적합한 문항으로 볼 수 있다. 2003년도 시험 결과를 BIGSTEPS로 분석하여 본 결과 적합도 통계치 (ZSTD) 50 문항 중 4문항이 -2.0 이하이거나 +2.0 이상이었다. 이 수치로 전반적으로 문항이 라쉬모형에 적합한 것으로 본다. 어느 정도 부적합 문항이 있어야 전체 문항이 적합한 것으로 보는 지는 특별한 기준이 없다. 문항의 적합도 정도가 꼭 문항반응이론에 필요한 전제라기보다는 적합하여야 문항반응이론의 도입이나 해석에 적절하기에 점검이 필요하다. 라쉬모형에서는 적합도는 주로 분별도의 적절성을 판단하는 자료로 삼는다.

이렇게 문항반응이론을 도입하여 분석할 수 있는 자료이므로, 그 후 동등화 작업을 알려진 방법의 절차를 따라 시행할 수 있었다. 고전검사이론에 따른 평균 및 표준편차 결과를 보면 2002년도 및 2003년도 수험생의 능력이 동등하다고 가정할 때, 2002년도의 문항이 더 어려운 문항이 많았다고 할 수 있다. 그러나 고려하여야 할 점으로 시험 치른 후 반드시 문제를 공개하는 것이 의과대학 시험의 원칙이고 다음 해의 문제에서는 과거 문항 중 지문이나 답가지를 바꾸어 같은 내용을 물어 보는 것이 있으므로 일부 문항의 노출 효과가 있다고 보아야 한다. 이런 동등화 시도는 한 과목에서의 연도별 시험 결과를 가지고 시행한 것인데, 몇 가지 기술적인 문제가 있다. 우선, 사전 설계된 동등화 작업이 아니므로 공통문항의 선별의 한계가 있다. 같은 내용을 가지

고 질문하는데 지문을 바꾼다거나 답가지를 바꾼다면 일정 능력의 수험생에서 난이도는 일정하다고 보기가 어렵기에 이 공통 문항을 선정하는 일이 가장 어려운 일이다. 만약 완전히 똑 같은 문제가 있고 그것이 전혀 노출되지 않은 상태에서 두 연도에 걸쳐 시험을 치른다면 문제가 없으나, 그런 상황은 거의 발생하지 않기에 완전히 똑같은 문제는 의과대학 시험에서는 대개 없다. 그러므로 이런 경우의 공통 문항은 완벽한 공통문항이라고 하기보다 가상 공통문항이다. 이 가상의 공통 문항일 경우에도 같은 내용을 물어 보는 문항을 고르고, 또한 그런 문항 중에서 난이도가 유사한 것을 공통 문항으로 설정하는 도리밖에 없다. 난이도가 유사한 것을 고를 때 가능하다면 거의 같은 난이도를 가진 문항이어야 하나, 그런 문항의 수가 지나치게 적어 내용이 공통인 문항 중 난이도 차이가 적은 것을 고를 때 임의로 난이도모수가 1 이하로 선정한 것은 특별한 기준으로 설정한 것은 아니고 차이가 적은 순으로 선정한 것이다. 이 공통 문항을 선정할 때 완벽하게 같은 문항이고, 시험 번호 위치도 같고 난이도 차이도 매우 적어야 한다는 등 엄격한 원칙에 따라야 한다고 하나 교육현장의 현실에서는 사전 설계를 정밀하게 하고 및 문항 노출을 막기 전에는 어려운 일이다. 이번에 시행한 검사 동등화는 처음부터 설계하여 시도한 것이 아닌, 두 해에 걸친 시험 성적에서 변이를 설명하기 위한 노력으로 시도한 것이기에 가상 공통문항이라는 한계가 있다.

이 연구에서 척도변환 후 분석 결과를 보면 문항 난이도 로짓의 척도값이 1 보다 크므로 난이도 값의 폭이 커지고, 문항의 난이도의 중앙값이 0.9630 이므로 이만큼 시험이 더 어려워졌음을 뜻한다. 즉 처음 문항분석에 나온 값을 보면 2003년도 문항이 2002년도 문항보다 어려운 것으로 나오는 데, 공통 문항을 보면 2003년도에서 더 난이도모수가 낮은(쉬운)문항으로 나오므로 그것을 보정하여 난이도 지수를 더 올려서 2002년도 문항과 동등화시킨다고 할 수 있다. 동등화 이후 2003년도 수험생의 능력모수가 더 커진 것으로 밝혀졌다. 그러나, 고전검사이론에서 2003년도 수험생이 더 시험 점수가 높은 것

은 상당 부분 문항의 노출에 기인한다고 볼 수 있으므로, 문항반응이론을 이용한 지금과 같은 분석법으로도 이런 문항 노출에 대한 보정은 어렵다. 동등화 전에도 2003년도 수험생의 능력모수가 높았음에 비추어 이런 능력 모수의 차이가 문항노출에 의한 것인지 아니면 참된 능력의 상승을 반영한 것인지 구별은 매우 어렵다. 단, 2003년도 수험생이 그 동안 어느 학년보다 수업태도 및 실습 태도, 수업 참여도가 뛰어났던 점은 있어, 그런 전반적인 학습 분위기의 호전이 능력이 뛰어난 집단에서 유래한 것이라 추측할 수 있으나 그 근거는 교원의 판단에 의존하는 것이라 객관 자료로 삼기는 어려운 문제이다. 과연 분석 결과처럼 2003년도 수험생이 2002년도에 수업을 들은 수험생보다 능력이 뛰어나다고 할 수 있는 지 이런 요인을 고려하여 언급하여야 한다.

이번 자료에서는 진점수 동등화나 관찰점수 동등화 결과에서 두 점수 중 어느 것을 사용하더라도 차이가 없었다 (Figs. 3, 4, 5). 그리고, 진점수와 능력모수 값은 선형 상관을 보이고 상관계수도 높았으나 능력모수의 상하 극단 값에서는 기울기가 급하였다 (Fig. 3). 극단값에서는 능력모수의 차이가 진점수에 영향을 덜 미침을 알 수 있다. 진점수 및 관찰점수의 상대 빈도는 2002년도 및 2003년도의 문항수가 차이가 나서 그림에서 빈도 분포가 다른 것으로 보이나, 문항수 차이를 감안하면 빈도가 거의 일치함을 알 수 있었다 (Fig. 6). 이 결과를 보면 진점수나 관찰점수 모두 추후 문항반응이론을 도입한 뒤, 성적을 보고할 경우 충분히 사용할 수 있음을 알 수 있다.

이번 동등화 작업에서는 우선 문항반응이론 중 라쉬모형을 도입하여 살펴보았으나 이렇게 수험생 수가 적을 때는 고전검사이론에 따른 동백분위동등화 (equipercentile equating)로 분석하여 보고 비교하여 보는 것이 다음의 과제이다. 이번 분석에서는 두 시험의 공통 문항의 난이도 상관관계에서 기울기가 1 에 가깝지 못하여 문항반응이론에 따른 동시 분석 (concurrent analysis)은 하지 못하였다. 만약 1 에 가까울 경우 이 두 시험 결과를 하나로 보고 분석하여 결과를 비교하여 볼 필요가 있다. 앞으로 다양한 방

법으로 계속 분석하여 보고 어느 방법이 가장 실용적인지 그리고 어떤 차이를 보이는 지 알아 볼 필요가 있다.

의과대학 시험의 동등화에 대한 과제로 우선 이 동등화를 어디에 사용할 것인지가 그 용도를 명확히 하여야 한다. 수험생의 능력의 변화를 볼 것인지 아니면 낙제 여부의 판단에 사용할 것인지, 학점을 줄 때 도입할 것인지 등을 고려할 수 있다. 진점수를 구하여 합격선 설정에 응용할 수 있고, 샘플이용 적응검사의 문제은행에 문항모수를 입력할 때 활용할 수 있다. 이 연구는 동등화 기법을 여러 풀그림을 가지고 시도하여 검사 동등화를 과목 단위에서 시행한 점에 의의가 있다.

### 참 고 문 헌

김수녕, 허선 (2002). 국산통계풀그림으로 우리 나라 의학 학술지에서 사용하는 통계를 해결할 수 있을까? *한국의학교육*, 14, 111-117.

지은림, 채선희(2000). 라쉬모형의 이론과 실제. 서울, 교육과학사.

남현우(2001). 검사 동등화 방법. 서울, 교육과학사.

BIGSTEPS [computer program] <http://www.winsteps.com>에서 받을 수 있음.

HRMH [computer program] Nandakumar 선생에게서 직접 받음.

Kolen MJ, Brennan RL(1995). Test equating: *Methods and Practices*. Springer, New York, U.S.A.

Nandakumar R(1993). A FORTRAN program for assessing essential unidimensionality data using Holland and Rosenbaum's methodology. *Multivariate Behavioral Research*, 28, 63-66.

PIE [computer program] version 1.0. Hanson B and Zeng L. American College Testing PO Box 168, Iowa City, Iowa 52243

ST [computer program] version 1.0. Hanson B and Zeng L. American College Testing PO Box 168, Iowa City, Iowa 52243

**부록 1. 자료 처리하기 위한 시험 결과 원자료**

2002년도 자료는 아래와 같다. 첫 열 및 둘째 열은 수험생의 번호를 표시하며, 셋째 열부터 맞으면 1, 틀리면 0 으로 변환한 자료로 각 문항은 열로 나타난다. 첫 줄은 1 번, 둘째 줄은 2 번 수험생의 반응을 뜻한다.

```
01100111111000110010101100110001000000100010001101001011100011
02100011110100000011101101111001000010111110100111001011111111
....
...
7400011010010000001111111111010111010101110001111100111110011
7510011111010000001111101011101110011101111111111110111101011
```

**부록 2. BIGSTEPS를 이용한 추정에 사용한 통제 파일 (control file) 명령어 (command) 예**

쌍반점 이후는 명령어에 대한 설명이다.

```
&INST
TITLE = "Rasch for 2002 exam"
PERSON = STDNT
ITEM = item
NI = 60 ; number of items
ITEM1 = 3 ; column of response to first item in data record
NAME1 = 1 ; 수험생 이름이 시작하는 열 번호
INUMB=Y ; 문항의 번호 표시
XWIDE = 1 ; number of columns per item response
CODES = 10 ; valid codes in data file
MISSING = 0 ; 반응하지 않은 문항은 틀린 것으로 친다.
DATA = para2002.dat ; 분석하는 원자료
IFILE = para2002.IF ; 문항모수 결과 파일이름
PFILE = para2002.PF ; 능력모수 결과 파일이름
CSV = Y ; 결과 파일을 csv 형태로 저장
TABLES = 11111111111111111111 ; output 파일에 나타나는 표로 1 이면 보이고 0 이
        면 보이지 않음
curves = 333 ; 결과 그림을 보임
&END
```

부록 3. ST의 통제 파일

Form X(2003, New Form), Form Y(2002,Old form)

\$ Item\_Parameters; 2003년도 시험 (Form X) 의 공통문항 (common items)의 문항모수를 뜻한다. 첫 열은 문항순서이고, 다음의 열인 분별도모수는 라쉬모형에서 모두 1로 설정하고, 다음 열인 난이도모수만 차이가 나며 마지막 열인 추측도모수는 모두 0이다.

1	1	-0.7	0
2	1	-1.03	0
3	1	-1.03	0
4	1	-1.23	0
5	1	-2.92	0
6	1	-4.13	0
7	1	-1.23	0
8	1	-1.03	0
9	1	-0.85	0
10	1	-0.85	0
11	1	0.07	0
12	1	0.07	0
13	1	0.53	0
14	1	0.98	0

\$ Link\_Item\_Parameters ; 2004년(Form Y)시험의 공통문항의 문항모수로 앞의 FormX 설명과 같다.

1	1	-0.06	0
2	1	-0.48	0
3	1	-0.56	0
4	1	-0.72	0
5	1	-2.12	0
6	1	-5.02	0
7	1	0.01	0
8	1	0.07	0
9	1	0.13	0
10	1	0.32	0
11	1	0.72	0
12	1	0.91	0
13	1	1.6	0
14	1	1.6	0

\$ Theta\_Distribution ; 2003년도 시험 (Form X)에서 모든 수험생의 능력 모수 나열

2.15  
.63  
.. Theta value of exam 2003  
2.67

\$ Link\_Theta\_Distribution ;2002년도 시험(Form Y)에서 모든 수험생의 능력모수 나열

-.58  
1.04  
... theta value of exam 2002  
.35  
1.16

부록 4. PIE의 통제 파일

---

```

$ Item_Parameters
1      1      -0.61    0
2      1      -0.92    0
..Item parameters of exam 2003 (New Form)
50     1      -0.13    0
$ Link_Item_Parameters
1      1      -0.72    0
2      1      3.77    0
.. Item parameters of exam 2002 (Old Form)
60     1      -3.1    0
$ Theta_Distribution
2.88
2.41
.. Theta distribution of exam 2003 (New Form)
4.03
$ Link_Conversion ; 첫 열은 2002년도 시험(Form Y)에서 원점수 0에서 60까지이므로
61개의 경우의 수가 있고, 다음 열에서 최저 점수는 0에서 시작하고, 세 번째 열에서
최대 점수는 60임을 뜻하며 네 번째 열은 점수가 1 점 단위 척도로 매겨짐을 뜻하며
다섯 번째 열은 척도 변환이 선형이면 1, 비선형이면 0인데 이 경우 선형이므로 1로
표기하였다.
Form Y
61     0      60      1      1
0      0      0      0
1      1      1      1
2      2      2      2
...
60     60     60     60
    
```

---