한국통계학회논문집 제1권1호
1994년 12월 pp. 206-214

대학별고사를 위한 문항분석, 표준점수, 검사동등화

성 태 제 1)

요 약

본 논문은 1994학년도 부터 부활된 대학별고사 실시에 따른 문항분석, 표준 점수제 그리고 검사동등화의 문제점을 지적하기 위하여 교육측정이론의 기본 개념을 소개하는데 있다.

대학별고사의 타당성과 신뢰성을 보장받기 위하여는 양질의 문항제작이 우선하여야 하며, 이를 위하여 문항분석은 종전에 사용하던 고전검사이론 보다는 문항반응이론을 이용하는 것이 바람직하다. 문항반응이론에 의한 문항분석은 피험자 집단의 특성에 의하여 문항특성이 달리 분석되지 않는 특징을 지니고 있기 때문이다. 문항이 논술형일 경우 채점자간 신뢰도와 채점자 내 신뢰도를 간과하여서는 않될 것이다.

다양한 선택과목을 채택하는 대학별 고사에서 입학 사정을 위하여 원점수를 사용하거나, 표준점수 혹은 검사동등화 방법을 이용하고 있으나 이는 교육측정이론에 위배된다. 다른 과목에 대한 인간의 능력을 상대비교 할 수 없으며, 표준점수와 검사동등화는 동일 능력에 대한 상대비교를 위한 방법이다. 특히 검사동등화는 동일 특성, 공정성, 모집단 불변성, 대칭성을 전제한다. 표준점수제에 의하여 수험생들의 다른 능력을 상대 비교하는 방법은 다른 능력이 점수로 표현되기 때문에 가능하나 그 점수가 무엇을 의미하는 가를 분석할 때는 교육평가의 기본 철학에도 위배된다.

Ⅰ. 서론

Thorndike(1918)는 존재하는 것은 모두 양으로 존재하기에 가시적이든 비가시적이든 측정할 수 있다고 하였다. 인간의 능력은 비가시적인 잠재적 특성(latent trait)이므로 검사라는 도구를 사용하여 측정할 수 있다. 측정은 정확성, 객관성, 효율성을 요구하기 때문에 구두시험, 필기시험, 질적 서열 부여, 양적 서열 부여, 논술형 검사, 단답형 검사, 이어 선다형 검사의 순으로 발전하여 왔다(Madus & Kellaghan, 1992). 이같은 측정의 역사적 발전은 인구의 증가로 상호경쟁에 의한 측정의 정확성이 요구되어 지면서 검사가 보다 객관화된 형태로 발전되었다 할 수 있다.

높은 교육열과 고등교육의 수요 공급의 차이로 홍익인간 구현이라는 명시적 교육 목표보다 상급학교 진학의 묵시적 교육목표가 강조되는 우리 현실에서 입시제도는 항상 사회의 관심이 되었으며 11번의 변화를 거듭하였다. 11번의 입시제도의 변화를 분석하면 대학의 질적 저하에 대한 우려와 대학 부정 입학을 방지하기 위하여 국가에서 주관하였다가, 시기적으로 대학의 자율권이 강조되면 대학에서 출제 관리하는 체제로의 순환의 연속이었으나 입시제도의 구성요소는 국가주도 검사, 대학자체 검사, 고등학교 내신성적으로 나뉜다(성태제, 1993a). 어느 제도하에서도 지필검사를 실시하지 않은 경우는 없었으며, 1994학년도 부터는 국가단위의 대학수학능력시험뿐 아니

1) (120-750) 서울시 서대문구 대현동 11-1 이화여대 사범대학 교육학과.

라 대학별고사를 실시하기에 이르렀다.

대학별고사는 대학본고사로 실시되어 오던 중, 소수과목 집중 고액과외의 사회적 폐단을 개선하기 위하여 1980년 7월 30일 발표된 과열과외 해소 방안으로 1981학년도 부터 중단되어 대학별고사 출제, 관리, 시행, 분석등의 발전 기회를 상실하였다가 1994학년도에 부활되었다. 이에 따라 1994학년도 대학별고사를 주관한 대학에서 시행착오와 문제점이 적지 않았을 것으로 추측된다.

본 논문은 대학별고사 실시에 따른 문항분석, 표준점수제, 그리고 검사동등화에 따른 문제점을 해결하기 위하여 교육측정의 기본이론을 소개하는 데 목적이 있다.

II. 검사의 타당도와 신뢰도

검사를 제작함에 있어 검사제작자와 문제출제진이 가장 먼저 고려해야 할 사항이 타당도와 신뢰도이다. 타당도는 검사 문항이 측정하고자 하는 내용을 측정하였느냐의 문제이며 측정 목적의 부합도로서 무게를 달기 위하여 저울을 사용하여야 한다는 뜻이다. 그러므로 대학별고사에서 국어 능력을 측정하기 위하여 국어 능력을 측정하는 검사가 제작되어야 한다. 타당도는 내용타당도, 구인타당도, 준거타당도로 구분되며 준거타당도는 공인타당도와 예측타당도로 나뉜다 (AERA/APA/NCME, 1985). 대학별 고사의 타당도를 검증하는 적절한 방법은 내용타당도와 예측타당도를 들수 있다.

내용타당도(content validity)는 내용전문가에 의하여 측정하고자 하는 것을 얼마나 잘 측정하는가에 대한 주관적 판단이다. 그러므로 내용타당도는 계량화하지 못하는 단점이 있으나 일차적으로 검증하는 타당도이다. 내용타당도를 보장받기 위하여 문항출제 전에 문항 출제진이 합의하여 이원분류표를 작성하여야 한다. 이원분류표에 의한 내용소와 행동소가 명확히 규정될 때 검사도구의 내용타당도는 보장받게 된다.

예측타당도(predictive validity)는 검사 점수가 피험자의 미래 행동이나 특성을 어느 정도 정확하게 예언하는냐의 문제로 대학별고사의 타당도를 검증할 수 있는 타당도의 한 방법이다. 즉 대학별고사의 점수가 입학 후 1학년, 혹은 2학년 나아가서 대학의 GPA를 얼마나 잘 예언하여 주느냐로 검사의 타당도를 검증할 수 있다. 이는 시간을 요하기 때문에 94학년도 대학별고사의 예측타당도는 최소한 1년 후에라야 검증할 수 있다.

신뢰도는 측정하고자 하는 것을 얼마나 일관성 있게 안정적으로 측정하였느냐의 문제이다. 일정하게 무게를 재는 저울이 신뢰도가 높은 저울이다. 신뢰도는 재검사 신뢰도, 동형검사 신뢰도, 내적일관성 신뢰도로 구분되며 내적일관성 신뢰도 안에 반분검사 신뢰도, KR-20, KR-21, Hoyt 신뢰도, Cronbach α가 있다. 내적일관성 신뢰도는 문항들이 얼마나 일관되게 하나의 특성을 측정하고 있는가를 말하여 준다. KR-20, Hoyt 신뢰도, 그리고 Cronbach α는 다음의 신뢰도 이론적 공식에 의하여 유도되었으므로 동일한 자료를 가지고 각기 다른 방법으로 신뢰도를 추정하여도 신뢰도계수가 같다.

$$\rho_{xx'} = \frac{\sigma_t^2}{\sigma_x^2} \ .$$

대학별고사의 신뢰도 추정방법은 KR-20, Hoyt 신뢰도 그리고 Cronbach α가 과학적이다. 일반적으로 SPSS 프로그램으로 Cronbach α를 추정하나, 검사의 문항 분석 뿐아니라 피험자

능력을 측정하는 많은 프로그램이 있으므로 이 프로그램들을 사용하면 동시에 많은 검사 정보를 얻을 수 있다. 이 프로그램으로 LOGIST(Wingersky, Barton, & Lord, 1982), BILOG(Mislevy, & Bock, 1984, 1986), ASCAL(Assessment System Corp., 1988), RASCAL(Wright, Mead, & Bell, 1979), ITEMAN(Assessment System Corp., 1988) 등이 있다.

고전검사이론에 의한 신뢰도의 측정오차 개념은 문항반응이론에서 정보함수 개념과 대응된다. Fisher가 정의한 정보함수는 정확성의 역의 개념에서 출발하였으며 문항정보함수와 검사정보함수가 있다. 신뢰도는 모든 피험자에게 동일하게 적용되나 문항정보함수와 검사정보함수는 피험자에 따라 다르다. 즉 능력이 다를 때 한 검사로 능력을 측정하였다면 능력의 측정오차는 피험자마다 다르다. 검사 도구가 피험자의 능력 수준과 동일할 때 피험자 능력을 보다 정확하게 측정하므로 측정오차가 작아 검사정보가 가장 높아지나, 검사 도구의 난이도가 피험자 능력수준과 멀리 떨어질 수록 측정오차는 커져서 검사정보가 낮아진다. 그러므로 피험자의 능력을 보다 정확하게 측정하고자 하면 피험자 능력과 유사한 난이도 수준의 문항으로 구성된 검사를 실시하여야 한다. 대학별고사를 위하여는 해당 대학에 지원하는 피험자 능력수준 범위에 해당하는 난이도 수준의 문항과 문항변별도가 높은 문항으로 구성된 검사를 제작하는 것이 바람직하다.

III.문항분석

검사도구의 타당도, 신뢰도, 정보함수를 높이기 위하여는 양질의 문항들이 제작되어야 한다. 검사의 구성요소는 문항이므로 문항분석은 검사도구의 질을 향상시키는 초석이 된다. 문항과 검사를 분석하는 검사이론은 고전검사이론과 문항반응이론으로 양분된다.

고전검사이론은 1900년대 초반 부터 이론적 발전을 거듭하여 1980년대 가서는 더 이상 발전되지 않는 이론으로 검사의 점수가 진점수와 오차점수로 구성되었다는 가정에서 전개된 검사이론이다.

$$X = T + E$$

또한 진짐점수는 측정이 불가능하므로 무한히 반복 측정하여 반복 측정치의 평균값으로 대치되는 가정을 전제로 발전되었다.

$$T = E(X)$$

문항반응이론은 총점을 진점수와 오차점수로 구분하는 것이 아니라 문항 하나하나에 의한 점수의 합으로 계산한다. 그러므로 문항반응이론은 문항특성곡선에 기초한다. 문항특성곡선이란 피험자의 능력 수준에 의하여 문항의 답을 맞힐 확률을 나타내며 [그림 1]과 같다.

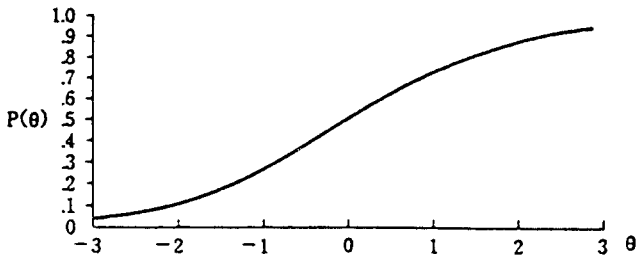

그림 1. 문항특성곡선

문항반응이론에 의하여 문항특성곡선을 나타내는 문항반응이론은 문항난이도를 고려하는 1-모수 모형, 문항난이도와 변별도를 고려하는 2-모수 모형, 문항난이도, 문항변별도, 그리고 문항추측도를 고려하는 3-모수 모형으로 구분되며, 정규오자이브 모형과 로지스틱 모형으로 나뉜다. 문항난이도와 문항변별도를 고려하는 2-모수 로지스틱 모형의 공식은 다음과 같다.

$$P_i(\theta_j) = \frac{1}{1+e^{-a(\theta-b)}}$$

a: 문항난이도, b: 문항변별도, θ: 피험자 능력수준.

2-모수 모형에서 문항난이도는 피험자가 문항의 답을 맞힐 확률이 .5에 해당하는 피험자의 능력수준을 말하며, 문항변별도는 문항특성곡선 상에서 문항난이도를 나타내는 점에서의 기울기를 말한다. 그러므로 문항특성곡선의 위치가 오른편으로 치우칠 수록 어려운 문항이 되며 기울기가 가파를 수록 변별력이 높은 문항이 된다. 문항반응이론의 기본개념 이해를 위하여 성 태제(1991), Baker(1985), Hambleton, Swaminathan, & Rogers(1991)의 저서를, 수리적 전개 이해를 위하여 Lord(1980)와 Baker(1992)의 저서를 참고하라.

문항분석은 고전검사이론과 문항반응이론에 의하여 모두 실시될 수 있다. 고전검사이론에 의한 문항난이도는 전체 피험자 중 문항의 답을 맞힌 피험자의 비율을 말하며, 문항의 변별도는 총점과 문항점수 간의 상관계수로 추정한다. 문항추측도는 답을 알지 못하고 추측하여 문항의 답을 맞힌 피험자의 비율을 말하며 계산공식은 다음과 같다.

$$P_G = \frac{W}{N(Q-1)}$$

W : 문항의 답이 틀린 피험자 수, Q : 보기수, N : 총 피험자 수.

문항의 추측도를 고려하여 교정한 난이도가 문항교정난이도이다. 문항교정난이도는 문항난이도에서 문항추측도를 뺀 값이 된다. 고전검사이론에 의한 문항분석에서 문항난이도는 피험자 능

력 집단의 특성에 의하여 모수의 추정치가 변화되고, 피험자 집단이 동질적인 경우 문항변별도는 낮게 추정된다. 또한 검사의 신뢰도와 측정오차를 추정할 때 모든 피험자에게 동일한 신뢰도와 측정오차가 적용되는데 비하여 실제 상황에서는 검사의 난이도 수준이 피험자 능력 수준과 유사할 때 신뢰도가 가장 높으며 측정오차가 가장 낮다. 이와 같이 문항 특성과 피험자 능력의 불변성 개념을 유지하지 못하는 고전검사이론의 문제점을 해결하기 위하여 제안된 이론이 문항반응이론이다.

Lawley(1943)가 수리적 모형을 전개하면서 발전된 문항반응이론은 일차원성과 지역독립성의 가정을 기초로 최대 우도추정법에 의하여 문항모수와 피험자의 능력모수를 추정한다. 문항반응형태에 의한 우도함수는 다음 공식과 같다.

$$L = \prod_{i=1}^{n} P_i(\theta_j)^{Ui} Q_i(\theta_j)^{1-Ui}$$

i: 문항, Ui: 정답여부.

우도함수에 의한 문항과 피험자 능력을 추정하는 수리적 방법은 결합 최대 우도추정법, 조건 최대 우도추정법, 주변 최대 우도추정법, 그리고 베이지안 추정법이 있다.추정된 문항난이도, 문항변별도 그리고 문항추측도는 b, a, c 혹은 β, α, c로 표기된다. 문항난이도 b는 주로 -2에서 +2 사이의 범위에 있으며, 문항변별도 a는 일반적으로 0보다 크고 1.5보다 작다. 문항추측도 c는 피험자 능력이 전혀 없는 피험자가 추측으로 문항의 답을 맞힐 확률로 일반적으로 사지 선다형 문항에서 .15 이하이다. 피험자의 능력은 θ로 표기하며 0에서 +∞의 범위를 지니나 문항반응이론에서 능력 척도에 의하면 -∞에서 +∞의 범위를 지닌다. 그러나 많은 피험자들의 능력 수준은 -3에서 +3에 존재한다.

문항반응이론이 고전검사이론에 비하여 수리적으로 복잡함에도 불구하고 문항분석이나 검사도구 개발을 위하여 널리 사용되고 있음은 불변성의 원리 때문이다. 문항반응이론은 피험자 집단의 능력에 관계없이 안정적으로 문항의 모수를 추정하는 문항특성의 불변성 개념과, 검사도구의 특성에 구애받지 않고 피험자의 능력을 일정하게 추정하는 피험자 능력 불변성 개념을 유지하고 있다. 그러므로 문항이나 피험자의 능력을 추정하기 위하여 문항반응이론에 기초한 LOGIST(Wingersky, Barton, & Lord, 1982), BILOG(Mislevy, & Bock, 1984, 1986), ASCAL(Assessment System Corp., 1988, RASCAL(Wright, Mead, & Bell, 1979), ITEMAN(Assessment System Corp., 1988) 프로그램 등을 사용하고 있다.

BILOG 프로그램은 주변 최대 우도추정법을 알고리즘으로 하여 짜여진 프로그램으로서 PC version이 개발되어 문항과 피험자 능력을 추정하는 보편적인 프로그램이 되었다. BILOG 프로그램은 문항반응이론에 의한 문항분석과 피험자 능력 추정 검사정보 추정 뿐 아니라 고전검사이론에 의한 문항분석과 신뢰도도 추정한다.

문항반응이론과 고전검사이론에 의한 양질의 문항이란 문항변별도가 높고 문항추측도가 낮은 문항이라 할 수 있다. 문항난이도에 대하여 좋은 문항인지 나쁜 문항인지를 평가하는 기준은 없다. 다만 규준참조평가의 목적을 지니고 있으면 다양한 수준의 난이도 문항이 필요하고, 준거참조평가의 목적을 지니고 있으면 준거에 대응되는 난이도 수준의 문항이 많이 제작되어야 한다.

양질의 검사도구를 제작하기 위하여 문항분석이외에 고려하여야 할 점이 차별기능 문항이다. 차별기능문항(differential item function)이란 능력 수준이 같음에도 불구하고 소속된 집단 특성

때문에 문항의 답을 맞힐 확률이 다른 문항을 말하며, 편파성 문항(biased item)이라고도 한다. 차별기능문항은 성별에 따라, 문화적 배경에 따라, 인종에 따라 나타난다. 1970년대 미국에서 지능 논쟁에 따른 후속 연구로 인종에 의한 차별기능문항이 연구 영역이 되었으며 국내에서의 관심은 초기 단계에 있다. 추정아와 성태제(1993a)는 대학수학능력시험 실험평가에서 성별에 따른 차별기능문항을 추출하고 있다. 자격증을 부여하든가 당락을 결정하는 행정적 의사 결정이 중요시 되는 모든 검사에서 차별기능문항은 제거되어야 한다. 차별기능문항을 추출하는 방법 중 최근에 널리 사용되고 있는 방법으로 Raju(1988, 1990)의 문항특성곡선간의 넓이 추정방법과 Holland와 Thayer(1988)의 Mantel-Haenszel 방법이 있다. 성 태제(1993b)는 Raju 방법과 Mantel-Haenszel 방법의 비교연구를 통하여 Raju 방법이 보다 정확하다 보고하고 있다. Raju 방법에 의하여 차별기능문항을 추출하는 프로그램으로 IRTDIF(Kim, & Cohen, 1991)가 있으며 Mantel-Haenszel 방법에 의하여 차별기능문항을 추출하는 프로그램으로 MHDIF(추정아, 성태제, 1993b)이 있다.

Ⅳ. 표준점수와 검사동등화

1994학년도의 대학별 고사에서 선택과목이 다양할 때 입학사정에서 크게 고려되어야 할 사항이 다른 교과목의 점수를 상호 비교하는 것이다. 즉 자연계열에서 선택과목으로 물리, 화학, 생물이 선정되었을 때 한 교과목에서의 몇점이 다른 교과목에서 몇점과 능력이 같은지를 분석하여야만 대학 입학사정이 공정하게 된다. 만약 획득점수를 그대로 입학사정에 반영한다면 보다 쉽게 출제된 과목을 선택한 학생은 유리할 것이고 어렵게 출제된 과목을 선택한 학생은 불리할 것이다. 이같은 문제점을 해결하기 위하여 표준점수나 검사동등화를 사용하고 있다. 표준점수에 의한 능력 비교나 검사동등화는 동일 특성을 측정하여 얻은 능력추정치에 의하여 실시되어야 한다. 대학수학능력시험과 같이 1차와 2차로 두 번 실시되었을 때 영역 별로 표준점수에 의한 비교와 검사동등화는 가능하다. 그러나 다른 영역에서 측정된 다른 특성들을 표준점수나 검사동등화에 의하여 비교한다면 이는 사과와 오렌지의 비교이다.

동일 특성을 측정한 후 피험자의 능력 비교를 위하여 흔히 표준점수를 이용한다. 표준점수는 피험자의 능력 분포가 정규분포임을 가정하여 해당 확률에 해당하는 상대적 비교점수를 의미한다. 교육측정에서 쓰이는 표준점수는 백분위 점수, z점수, t점수, Stanine점수가 있으며 SAT, GRE, ACT점수 그리고 지능점수로 Wechsler점수와 Stanford-Binet점수도 표준점수이다(Hopkins, Stanley & Hopkins, 1990).

검사동등화는 한 검사에서 얻은 점수를 다른 검사에서 얻은 점수로 환산할 수 있도록 두 검사 점수들간의 규칙적 관계를 찾아내는 과정으로, 동등화를 위한 조건은 동일 특성(same ability), 공정성(equality), 모집단 불변성(population invariance) , 대칭성(symmetry)이 유지되어야 한다(교육평가연구회, 인쇄중; Angoff, 1971; Lord, 1980).

동등화의 형태는 수평동등화와 수직동등화로 구분된다. 수평동등화(horizontal equating)란 두 개의 동형검사로 유사한 능력 집단의 피험자에게 검사를 각각 실시한 후 동등화하는 형태이며 수직동등화(vertical equating)란 검사의 난이도 수준도 다르고 피험자의 능력 수준도 다를 때 사용하는 동등화 형태이다. 2차이상 실시한 대학수학능력시험의 경우는 수평동등화의 형태이다.

동등화를 위한 응답자료 수집을 위한 동등화 설계는 단일집단 설계, 동등집단 설계, 가교검사 설계로 나뉜다. 단일집단 설계(single-group design)는 한 집단이 다른 두 개의 검사를 치루었을 때 동등화를 위한 설계이며, 동등집단 설계(equivalent group design)는 능력 수준이 유사한 두 피험자 집단이 동형검사를 치루었을 때 실시하는 설계이다. 가교검사 설계(anchor-test design)란 동등화될 검사들을 어떤 형태로 연결시키는 공통문항을 사용하는 방법으로 외적가교검사(external anchor test), 내적가교검사(internal anchor test) 그리고 통합절차(pooled procedure)가 있다.

동등화를 위한 수리적 방법은 크게 고전검사이론과 문항반응이론에 의하여 구분되며 고전검사이론으로 선형동등화와 백분위 동등화법이 있다. 문항반응이론에 의한 방법은 평균과 표준편차방법(Loyd & Hoover, 1980; Linn, Levine, Hasting & Wardrop, 1980), 회귀분석방법, 검사특성곡선 방법(Stocking, & Lord, 1983)이 있다. Baker, Ali Al-Karni(1991a)는 동등화 방법 중 문항반응이론에 기초한 검사특성곡선 방법이 우월하다고 주장하였다. 검사동등화의 형태, 설계, 방법은 검사의 상황과 특성, 피험자 집단의 특성에 따라 각기 다양한 유형으로 동등화되나 일반적으로 문항반응이론에 의한 검사동등화 방법이 선호되고 있다. 국내 연구는 남 현우(1992), 황소림(1993)의 연구가 있으며 Holland와 Rubin(1984)이 편집한 저서를 참조하라. 검사동등화를 위한 프로그램으로 검사특성곡선방법에 의한 EQUATE(Baker, Ali Al-Karni, 1991b)가 있다.

문항분석 뿐 아니라 검사동등화에 있어 수리적으로 어렵고 힘겨운 방법이 제안되고 있는 것은 측정의 정밀성(precision of measurement)을 추구하기 때문이다. 측정의 정밀성에 앞서 기본 전제가 되는 것은 측정의 내용이며, 측정이론의 적용을 위한 기본 가정이 충족되느냐 하는 문제이다.

대학별 고사가 다양한 선택과목을 설정하였을 때 능력 수준 비교를 위하여 선택과목간의 표준점수나 검사동등화를 실시하는 경우가 있다. 그것은 점수가 수이기에 가능하다. 그러나 수의 내용이 무엇인가를 생각할 때 비교가 가능한지는 의문이 생기게 된다. 또한 각기 다른 과목을 선택한 피험자를 동일한 척도에 의하여 평가하는 것은 교육평가이론에도 위배된다.

참고문헌

[1] 교육평가연구회(인쇄중). 「교육측정, 평가, 연구, 통계 용어사전」. 서울:중앙교육진흥연구소.

[2] 남 현우(1992). 문항모수 변이에 따른 선형, 동백분위, IRT, 검사동등화 방법의 강인성 비교연구. 「교육평가연구」, 제 5권 제 2호, 27-60.

[3] 성 태제(1991). 「문항반응이론 입문」. 서울:양서원.

[4] 성 태제(1993a). 입시위주의 교육과 과열과외. 「교육학연구」, 제 31권 제 2호, 67-86.

[5] 성 태제(1993b). 차별기능(편파성)문항 추출을 위한 Raju 방법과 Mantel-Haenszel 방법 의 비교연구, 「교육평가연구」, 제 6권 제 1호. 91-120.

[6] 추정아, 성태제(1993a). Mantel-Haenszel 방법과 Raju 방법에 의한 제 4차, 제 5차 대학수학능력시험 실험평가의 성별에 따른 차별기능문항 추출. 「교육평가연구」, 제 6권 제 2호, 259-286.

[7] 추정아, 성태제(1993b). 「MHDIF: Mantel-Haenszel 방법을 이용한 차별기능문항 추출 컴퓨

터 프로그램」. 서울:이화여자대학교.
[8] 황소림(1993). 대학수학능력시험 제6차, 제7차 실험평가의 문항특성과 피험자능력 동등화. 「교육평가연구」, 제 6권 제 2호, 287-314.
[9] AERA/APA/NCME(1985). *Standards for educational and psychological testing.* Washington, DC:APA
[10] Angoff,W.H.(1971). Scales, Norms, and Equivalent Scores. In R.L.Thorndike(2nd. Ed.), *Educational Measurement, 508-600.* Washington, DC: American Council Education.
[11] Assessment Systems Corporation(1988). *User's manual for the MicroCAT testing system(Version 3).* St. Paul, MN: Author.
[12] Baker,F.B.(1985). *The Basic of Item Response Theory.* Portsmouth, NH:Heinemann.
[13] Baker,F.B.(1992). *Item Reponse Theory: Parameter estimation techniques.* New York:Marcel Dekker,Inc.
[14] Baker,F.B., & Ali Al-Karni(1991a). A comparision of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*,147-162.
[15] Baker,F.B., & Ali Al-Karni(1991b). *EQUATE: Computer program for equating two metrics in item response theory [Computer Program].* Maidison WI: University of Wisconsin, Laboratory of Experimental Design.
[16] Hambleton,R.K., Swaminathan,H., & Rogers,H.J.(1991). *Fundamentals of Item Response Theory.* Newbury Park: Sage Publications.
[17] Holland,P.W., & Rubin,D.B.(Eds.).(1982). *Test Equating.* New York: Academic Press.
[18] Holland,P.W., & Thayer,D.T.(1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H.I.Braun(Eds.), *Test validity, 129-146.* NJ.: Lawrence Erlbaum Associates.
[19] Hopkins,K.D., Stanley,J.C., & Hopkins,B.R.(1990). *Educational and Psychological Measurement and Evaluation(7th ed.), 55-62.* Englewood Cliff, New Jersey:Prentice Hall.
[20] Kim,S., & Cohen,A.S.(1991). *IRTDIF:A computer program for the IRT differential item functions [Computer Proram].* Madison WI: University of Wisconsin-Madison.
[21] Lawley,D.N.(1943). On problems connected with item selection and test construction. *Processings of the Royal Society of Edinburgh, 18,* 1-11.
[22] Linn,R.L., Levine,M.V., Hasting,C.N., & Wardrop,J.L.(1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159-173.
[23] Lord, F. M.(1980). *Applications of item response theory to practical testing problems, 193-211.* Hillsdale, NJ: Erlbaum.
[24] Loyd,B.H. and Hoover,H.D.(1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.
[25] Madus,F.G., & Kellaghan,T.(1992). Cruuiculum evaluation and assessment. In P.W.Jackson(Ed.), *Handbook of research on curriculum, 119-154.* NY: Macmillan Publishing Company.
[26] Mislevy,R.J., & Bock,R.D.(1984). *BILOG: Maximum likelihood item analysis and test*

scoring with logistic models. Mooresville, IN:Scientific Software.

[27] Mislevy,R.J., & Bock,R.D.(1986). *PC-BILOG: Item analysis and test scoring with binary logistic model [Computer Program]*. Mooresville IN: Scientific Software.

[28] Raju,N.S.(1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

[29] Raju,N.S.(1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

[30] Stoking,M.L. and Lord,F.M.(1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

[31] Thorndike,E.L.(1918). The nature, purposes and general methods of measurements of eudcational products. In G.M. Whipple(Ed.), *The Measurement of Educational Products, Seventeenth Yearbook of the National Society for the Study of Eduction, Part II*, 16-24. Bloomington, IL:Public School Co.

[32] Wright,B.D., Mead,R.J., & Bell,S.R.(1979). *BICAL: Calibrating items with the Rasch model*(Statistical Laboratory Research Memorandum No.23B). Chicago: University of Chicago, School of Education.

[33] Wingersky, M.S., Barton, M.A., & Lord,F.M.(1982). *LOGIST user's guide.* Princeton, NJ:Educational Testing Service.