

## 문항분석 이론에 의한 중학교 수확평가 학군간 차이 비교 연구

정 선 영 (언북중학교)  
임 형 (국립교육평가원)  
이 영 하 (이화여대)

### I. 서론

#### A. 연구의 필요성

대학입시에 내신성적 반영율이 40%로 크게 늘어났다. 그 이유는 무엇일까? 그것은 내신제도 만큼 교육적 효과가 크고 교육의 정상적 운영에 도움을 주는 제도도 드물기 때문이다.

교육적 효과의 극대화와 교육의 정상적 운영은 모두가 바라는 바이고 해결해야 할 과제임이 틀림없다. 그렇다면 내신제도는 이것에 어떻게 기여할 것인가? 즉, 내신제도를 실시함으로써 개선될 현재의 문제점은 무엇인가?

중학교를 중심으로 살펴보면, 가장 먼저 학원과외를 들 수 있다. 많은 수의 학생들이 방과 후 영어 수학 학원에 다니고 있다. 그 중에는 학교 수업 내용을 충분히 이해하고 좀 더 심화된 학습을 위해 학원 수업을 하는 학생들도 있다. 그러나 문제는, 수업 내용의 이해없이 그저 남들 따라하기 식의 태도로 학원을 다니는 학생들이다. 학원은 대체로 학교보다 어려운 문제를 다루고, 많은 양의 문제를 다루기 때문에 그들에겐 과중한 학습량이다. 학교 수업을 충분히 이해하지 못한 학생들이 과외 내용까지 소화해 낼 수 있을까? 게다가 이 학생들은 학교수업에 집중하지 않는다. 학원에서 똑같은 내용을 했거나 할 것이기 때문이다. 이것이야말로 시간의 낭비이다. 이 학생들을 수업으로 끌어들이야 할 것이다.

둘째, 현재 인문계 고등학교 진학을 위해 고입 선발고사를 실시하고 있는데, 이것은 인적 물적 자원의 낭비이다. 1991년의 경우 중학교 졸업자의 94.0%가 고등학교에 진학하였다.(한국의 교육지표; 1991). 졸업자의 대부분이 진학을 하고 있는 상황에서 따로 선발고사를 치른다는 것은 낭비일 것이다.

세째, 공신력이 있다는 장점 때문에 선발고사를 택하고는 있으나, 3년간 학생들의 학업성취도를 측정한 모든 점수들은 반영이 되어야 한다. 따라서 선발고사보다는 내신 성적을 학업성취도 기준치에 따라 절대평가하는 방안이 합당하다. 즉, 공교육에 믿음을 갖게 하는 것이 무엇보다 중요하다.

이처럼 내신성적의 반영은 꼭 필요한 일이지만, 실행시 야기되는 문제점 또한 사회적으로 논의의 대상이 되고 있다.

3월21일 金淑喜 교육부 장관이 참석한 전국 15개 시도 교육감 회의에서 현행 내신 제도의 문제점이 논의되었다. 1980년 7월30일 "교육정상화 및 과열과외 해소방안"이라는 교육개혁조치로 처음 시행된 고교 내신제가 15년동안 지역간 학교간 편차를 무시한 채 일률적으로 적용돼 많은 문제점을 노출하고 있기 때문이다.

이러한 내신성적의 일률적 반영은 고교 교육과 일선상에 있는 중학교 교육에 당연히 영향을 주었다. 지난해 12월 94학년도 고입 선발고사 때, 내신성적의 불리함을 이유로 비평준화 지역의 옛 명문고를 기피하여 하위권 고교로

학생들이 물리는 바람에 옛 명문고는 미달사태를 빚고 하위권 고교는 정원초과로 불합격자가 생기는 현상이 발생했다. 내신등급을 잘 받기 위해서 하향 지원하여 안주하기보다는 능력껏 도전하고 발전하는 삶에서 보람을 찾도록 가르침이 마땅하나, 이런 현상이 일어날 수 밖에 없는 이유는 각 학교의 내신성적이 차이가 있다는 생각 때문이다.

도시 학교와 농촌학교간, 공립학교와 사립학교 간, 심지어는 도시내의 학교간에도 내신성적의 차이가 있다. 그렇다면 실제로 그 차이는 무엇인지, 또 얼마나 차이가 나는 지 연구해 볼 필요가 있다.

### B. 연구의 목적과 연구문제

현재 중학교에서 실시되고 있는 수학시험의 양호도와 질적인 차이를 비교하여, 내신성적의 일괄적인 반응이 합당치 못함을 밝히는 것이 본 연구의 목적이다. 특히 학군간의 차이에 그 초점을 두고 있다.

연구문제는 다음과 같다.

1. 고전검사이론을 바탕으로 문항의 난이도, 변별도를 구하고 각 시험지의 신뢰도, 답지반응 분포에 따른 오답지 매력도를 구한 후 학교간 차이를 비교한다.
2. 문항반응이론을 바탕으로 문항의 무수치를 추정하여 학교간 차이를 비교한다.
3. 문항반응이론을 바탕으로 피험자의 능력 모수치를 추정하고 각 시험지의 정보함수를 찾아 피험자의 능력수준이 얼마나 정확하게 추정되고 있는지 학교간 차이를 비교한다.

### C. 연구의 제한점

연구대상이 서울시내 세 학교만으로 구성되었기 때문에 연구결과의 일반적 해석은 곤란하며, 특히 학군 별로 선정된 세 학교는 시험범위가 같은 학교로 선정해야 했기 때문에 각 학교

가 학군의 특성을 얼마나 대표할 수 있는 지 문제가 있을 수 있다.

## II. 교육 평가

### A. 교육평가의 개념

모든 교육과정과 교육활동은 반드시 우리가 성취하기를 기대하고 이상으로 하는 교육목표를 내포한다. 이러한 기대와 이상이 현실적인 교육활동을 통해서 얼마만큼 그리고 어느 정도까지 달성되었는가를 재어보고 판단하는 작용이 교육평가이며, 인간이 갖고 있는 개인차를 현상 그대로 정확히 밝혀주고 개인차를 차츰 좁혀서 0으로 떨어뜨리는 길을 모색하는 것이 교육평가의 목적이다.

학교현장에서 평가의 모든 영역처럼 보편화된 학생평가는 학생 개인의 학업성적의 진보와 결과에만 관심을 갖는 평가이다. 반면에, 교육과정 평가는 학교에서 투입하는 교육과정 및 프로그램 자체의 평가에 관심을 두고, 환경의 평가, 학습자의 평가, 학습의 평가, 교수방법의 평가등을 포함하고 있는 학생평가 보다 좀 더 포괄적인 의미의 평가라고 할 수 있다.

교육평가를 실시함에 있어 어떤 철학적 관점을 지니느냐에 따라 규준지향평가(norm-referenced evaluation)와 준거지향평가(criterion-referenced evaluation)로 구분한다. 규준지향평가(상대평가)는 개인의 성취수준 및 측정치를 비교집단의 규준에 비취 상대적 서열에 의하여 판단하는 평가이다. 즉 학생이 무엇을 얼마만큼 알고 있느냐에 주안점을 두기보다는 그 학생이 다른 학생과 비교하여 어느 지점에 위치하느냐에 관심을 두고 있다. 규준지향 평가는 일반적으로 총합평가시 사용되며 책무성을 부여하거나 제한된 인원을 선발하는 데 사용된다.

준거지향평가(절대평가)는 개인의 상대적 비교보다는 무엇을 얼마만큼 알고있느냐를 판단하는 평가로서 어떤 목표 혹은 준거에 비추어

개인의 학업성취 수준이 지식획득의 연속성 상에서 어느 지점에 있는가를 알고자 한다.

### B. 평가의 측정학적 특징

좋은 평가도구가 되기 위한 조건에는 타당도(validity), 신뢰도(reliability), 객관도(objectivity)가 있다.

타당도는 첫째, 무엇을 측정하고 있는냐, 둘째, 그 측정하려는 것을 어느 정도로 충실히 측정하고 있는냐의 문제이며, 무엇에 비추어 본 타당도인지 반드시 준거의 개념이 수반된다.

타당도를 분류하면 첫째, 목표 타당도(criterion validity)이다. 내용 타당도(content validity)는 평가하려고 하는 내용을 평가도구가 어느 정도로 충실히 측정하고 있는지를 분석하려는 타당도인 반면, 목표 타당도는 문항이 얼마나 성취 불성취를 예리하게 판단할 수 있는가에 대한 타당도이다. 이를 위한 방법으로는, 내용 전문가간 합치도 지수로 계산하는 방법이 있고, 다른 하나는 피험자 반응에 기초하여 각 문항에 대한 기대 난이도 수준과 관찰 난이도 수준을 비교하는 방법이 있다.

둘째, 예언 타당도이다. 평가도구가 목적하는 미래의 행동 특성을 정확히 높게 예언하는 정도를 말한다. 즉, 선행검사 X와 미래의 행동특성이라는 준거 Y와의 상관계수로 표시한다. 따라서 예언 타당도는 검사 X와 준거 Y의 두 평가도구의 신뢰도에 크게 영향 받는다.

셋째, 공인 타당도이다. 예언 타당도와 마찬가지로 검사 X와 준거 Y사이의 상관계수로 나타낸다. 그러나 차이점은 시간차원과 준거의 성질이 다르다는 것이다. 공인 타당도는 검사 X와 행동 준거 Y와의 사이에 '현재' 어느 정도로 잘 일치하느냐로 판단하는 것이고, 준거의 성질도 예언이 아니라 공통된 요인이 있는냐 없느냐에 있다. 즉, 검사 X로 검사 Y를 대체할 수 있는냐의 문제가 공인 타당도이다.

네째, 구인 타당도이다. 구인(construct)이란

검사성취에 반영되어 있다고 짐작되는 인간의 어떤 가정적 속성을 말한다. '이 특성을 가진 사람은 X라는 상황 아래서는 Y의 행동을 보일 것이다.'라는 법칙 정립이 구인 타당도의 핵심이다. 따라서 검사 자체 보다는 준거의 확인과 이론형성이 보다 중심이 되는 타당화 방법이다.

신뢰도는 측정하려는 것을 얼마나 정확하게 얼마나 오차없이 측정하고 있는냐의 개념이다. 즉, 측정하고 있는 정도에 일관성(consistency)이 있는냐, 측정의 오차(error of measurement)가 얼마나 적으냐 하는 질문이다. 신뢰도가 낮으면 타당도는 이에 비례해서 낮아지기 때문에 신뢰도는 타당도의 선행조건이 된다. 즉 신뢰도 없이 타당도가 높은 평가도구는 존재할 수 없다.

신뢰도를 추정하는 방법은 첫째, 재검사 신뢰도이다. 한 개의 검사를 같은 집단에게 두 번 실시해서 그 전후의 결과에서 얻은 점수를 기초로 해서 상관계수를 산출한다.

둘째, 동형검사 신뢰도(equivalent-form reliability)이다. 미리 두 개의 동형 검사를 제작하고 그것을 같은 피험자에게 실시해서 두 동형 검사에서 얻은 점수 사이의 상관을 산출한다. 이것을 동형성 계수(coefficient of equivalence)라고 한다.

셋째, 반분 신뢰도(split-half reliability)이다. 한 개의 검사를 한 피험집단에게 실시한 다음 그것을 적절한 방법에 의해 두 부분의 점수로 분할하고 이 분할된 두 부분을 독립된 검사로 생각해서 그 사이의 상관을 계산한다. 이것은 동질성 계수(coefficient of homogeneity)를 보는 한 가지 방법이다.

네째, 문항 내적합치도(inter-item consistency)이다. 검사 속의 각 문항을 독립된 한 개의 검사 단위로 생각하고 그 합치성, 동질성, 일치성을 종합한다. 만약 어떤 검사가 동형 검사 신뢰도 및 반분 신뢰도에서 높은 계수를 보이고 문항 내적합치도(Kuder-Richardson 계수)에서 낮은 계수를 보였다면, 그것은 검사 속에 표본

되어 있는 문항 사이의 동질도가 낮기 때문이라고 추정한다. 따라서, 반분신뢰도와 Kuder-Richardson계수의 차는 한 검사의 동질도지수로 사용한다.

Kuder-Richardson의 신뢰도 추정방법의 계산은 K-R 20 과 K-R 21 이 있다.

$$\text{K-R 20} : r_{xx} = \frac{n}{n-1} \left[ 1 - \sum \frac{pq}{s_x^2} \right]$$

$$\text{K-R 21} : r_{xx} = \frac{n}{n-1} \left[ 1 - \frac{x(n-x)}{ns_x^2} \right]$$

n: 검사 속의 문항의 수

p: 각 문항에 정답을 한 학생의 비율

q: 각 문항에 오답을 한 학생의 비율:

$$q=1-p$$

$s_x^2$ : 전체 검사점수의 변량: 검사점수의 표준오차의 제곱

x: 전체 검사점수의 평균

다섯째, 가장 측정학적 결함이 적은 Cronbach의  $\alpha$  계수이다.

$$\alpha_K = \frac{n}{n-1} \left[ 1 - \sum \frac{s_i^2}{s_x^2} \right]$$

$s_i^2$ : 각 단일 문항의 변량

$s_x^2$ : 전체 검사점수의 변량

n: 검사 속의 문항의 수

만약 문항을 0, 1 즉 <틀렸다>, <맞았다>는 두 가지 방법으로 채점하는 경우에 위 식은 K-R 20을 구하는 식으로 변한다.

객관도는 채점자의 채점이 어느 정도 신뢰롭고 일관성이 있느냐로서, 여기서는 큰 문제가 되지 않는다.

### C. 평가의 절차와 계획

평가의 절차와 계획의 첫 단계는 교육목적의 설정과 분석이다. 교육 목적을 설정하고 분석해서 그 개념을 정확하게 규정, 확인하는 것이 평가에서의 제일 먼저 할 일이다. 교육목적은 교

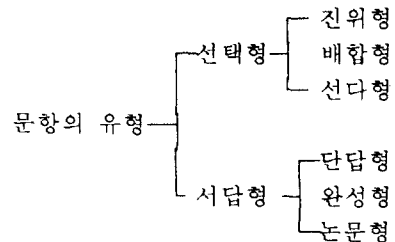
육의 모든 작용에 방향감을 주는 나침반의 역할을 하기 때문이다. 교육목적은 여러 수준의 것으로 생각해 볼 수 있으며, 상위의 교육목적은 추상적이고 막연하며 선택적인 진술보다는 포괄적인 진술로 되어있고 하위의 교육목적일수록 보다 구체성을 띤다. 교육목적은 각각의 하위단계를 거쳐서 궁극적으로 단원의 목적에 반영되고 직결되어야 하며 측정하려는 각 영역의 교육목적과 내용을 분명히 의식하고 진술해 두는 것이 필요하다.

두번째 단계는 평가도구의 제작이다. 먼저 교육목적을 행동면과 내용면으로 분류하고 이것의 이원목적분류표를 작성한다. 내용분류란 학생이 학습한 내용을 적절한 기준에 의하여 분류하는 일을 말하는데 일반적으로 각 교과와 단원이나 내용을 말하며, 행동영역은 구성하고 있는 교육목표에 따라 지적 영역(cognitive domain)과 정의적 영역(affective domain) 그리고 운동 기능적 영역(psych-motor domain)으로 구분된다. 이들 각각의 영역을 세분화 하면, 지적 영역; 지식, 이해력, 적용력, 분석력, 종합력, 평가력

정의적 영역; 감수반응, 가치화, 조직화, 인격화

운동 기능적 영역; 관찰, 모방, 연습, 적용 등으로 나눈다.

이원목적분류표가 작성되면, 분류표를 바탕으로 문항을 작성한다. 문항의 유형은 다음과 같다.



세번째 단계는 평가 실시 및 결과처리이다. 학생들에게 평가를 실시하고 결과를 바탕으로 검사문항의 질과 양호도를 분석한다. 현재 학교

에서는 각 과목의 평균과 석차를 알아보는 것으로 결과 처리를 끝내지만, 문항분석에 더 관심을 기울여서 학습지도에 활용해야 할 것이다.

문항분석을 위한 이론에는 검사의 총점에 의하여 문항을 분석하는 고전검사 이론과 문항을 각각의 고유한 문항특성에 의해 분석하는 문항 반응 이론이 있다.

### III. 문항분석 이론

#### A. 고전검사 이론(classical test theory)

고전검사 이론은 구조가 단순하고 분석결과의 해석이 용이하므로 대부분의 검사에 적용할 수 있는 유용성이 있다. 그러나 문항이나 검사 통계치가 피험자의 집단에 따라 다르게 추정된다는 약점이 있고, 피험자의 능력을 추정할 때 점수오차를 동일하게 적용한다는 적절치 못한 개념을 갖고 있다.

##### 1. 문항 난이도

문항 난이도는 문항이 얼마나 어려운가의 정도를 말하며 문항 j에 정답을 한 학생의 비율  $P_j$ 로 정의한다. 문항 난이도는

$$P_j = \frac{r_j}{n} \times 100$$

$r_j$  ; 문항의 정답자 수

$n$  ; 전체 피험자 수

위 식에서, 추측 요인이 강하게 작용했을 때 난이도는 과대 추정된다. 추측에 의한 정답자들을 정답자 수에서 제외하면  $P'_j$ 는

$$P'_j = \frac{r_j - \frac{w_j}{k_{j-1}}}{n} \times 100$$

$k_j$  ; 문항 j의 답지 수

$w_j$  ; 문항 j의 오답지 수

$n$  ; 전체 피험자 수

일반적으로 규준지향검사는 난이도가 20-80%인 문항으로 구성되며 20-80% 사이에서 정규분포를 하도록 하는 것이 적절하다. 즉 문항의 난이도가 50%인 것이 가장 많고 차츰 적은 수의 문항이 20-80% 수준의 난이도를 가지고 평균 난이도가 50%가 되도록 하는 것이 바람직한 것으로 알려져 있다.

준거지향검사일 때는 시험의 목적에 따라 문항의 난이도를 매우 낮게 또는 높게 정할 수 있다.

##### 2. 문항 변별도

문항 변별도는 각 문항이 검사가 측정하려는 특성 또는 능력을 가지고 있는 사람과 그렇지 않은 사람을 변별하는 정도를 말한다.

변별도를 개념화 하는 방법은 크게 두가지로 나누어 진다. 하나는 상위집단과 하위집단의 난이도의 차이로 구하는 방법이고 다른 하나는 각 문항의 점수와 전체점수의 상관계수에 의해 구하는 방법으로 양분점 상관계수(point biserial correlation coefficient)와 양분 상관 계수(biserial correlation coefficient)로 계산한다. 양분 상관계수는 한 문항에 정답과 오답을 하는 능력은 연속이라는 가정하에서 문항점수와 검사점수와의 상관이다.  $r_{pb}$  는

$$r_{pb} = \frac{(x_j - \bar{x})}{s_x} \times \left( \frac{P_j - P_j'}{t} \right)$$

$\bar{x}$  ; 전체 피험자의 평균점수

$x_j$  ; 문항 j에 정답한 학생들의 평균점수

$t$  ; 표준 정규분포에서  $P_j$ 에 해당하는 z 점수의 표준오차

$s_x$  ; 검사점수의 표준오차

$P_j$  ; 문항 j의 난이도

$P_j'$  ;  $1 - P_j$

한 문항에 정답과 오답을 하는 능력은 연속 변수이므로 양분 상관계수의 활용이 보다 바람직 하다. 상관계수의 언어적 해석을 제시하면 다음과 같다.

<표 3-1> 상관계수의 언어적 해석

.90 - 1.00	아주 상관이 높다
.70 - .90	상관이 높다
.40 - .70	확실히 상관이 있다
.20 - .40	상관이 있기는 하나 낮다
.00 - .19	상관이 거의 없다

3. 신뢰도

문항의 난이도  $p_j$ 와 변별도  $r_{jk}$ 를 사용하여 검사의 신뢰도를 계산할 수 있다. n개의 문항으로 구성된 검사의 평균점수  $\bar{x}$ 와 표준편차  $s_x$ 는

$$X = \sum_j p_j$$

$$S_x = \sum_j s_j r_{jk} = \sum (p_j q_j)^{1/2} r_{jk}$$

검사의 신뢰도 계수(reliability index)  $r_{xx}$ 는 다음과 같이 계산된다.

$$r_{xx} = \frac{n}{n-1} [1 - \sum_j \frac{s_j^2}{\sum (s_j r_{jk})^2}]$$

4. 답지반응 분포와 오답지 매력도

문항의 답지에 대한 반응의 분포상태가 답지 반응분포이다. 각 답지의 반응분포를 분석하여 오답지의 능률성을 조사하면 보다 매력적이고 제 능력을 가질 수 있는 오답지를 만드는 데 도움이 된다. 오답지가 제 구실을 하려면 매력 있어야 하며 오답지가 함정이 되어서는 안된다. 그러기 위해서는 모든 오답지가 매력 있고 고른 답지 반응분포를 나타내는 것이 좋다. 4지택 1형 문항의 매력도가 어느 정도인 지를 직관적으로 파악하기 위해서 다음과 같은 식을 이용한다. 오답지의 매력도  $AVI_i$  (%)는

$$AVI_i(\%) = \frac{f_i}{\frac{w}{k-1}} \times 100$$

$f_i$ : 오답지의 응답자 수

$k$ : 답지의 수

$$w = \sum f_i \quad i=1, 2, 3$$

위 식은 i 번 오답지의 실제 응답자 수 ( $f_i$ )를 한 오답지에 해당될 평균 응답자 수 ( $w/3$ )로 나눈 비의 백분율 의미한다.

오답지 매력도 지수의 언어적 해석은 다음과 같다.

<표 3-2> 오답지 매력도지수의 언어적 해석

오답지 매력도	언어적 해석
50% 이상	양 호
40% - 49%	판 단 유 보
39% 이하	불 량

B. 문항 반응이론

1. 문항 반응이론의 개념과 기본가정

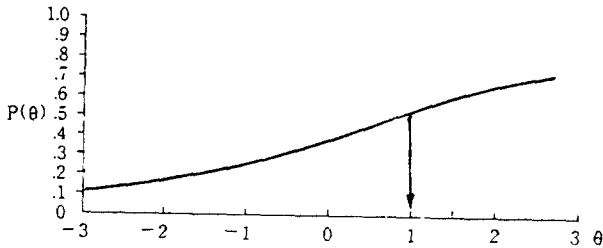
문항 반응이론은 관찰 가능한 피험자의 검사 결과와 그 결과에 잠재되어 있는 특성 또는 능력과의 관계에 관한 이론이며 불변성 개념(invariance concept)을 가지기 때문에 고전검사 이론에 비하여 논리적으로 타당하다.

불변성 개념에는 문항특성의 불변성 개념과 피험자 능력의 불변성 개념이 있다. 문항특성의 불변성 개념은 문항마다 고유한 특성이 있기 때문에 피험자 집단의 특성에 의해 문항의 특성이 다르게 추정되지 않는다는 개념이고, 피험자 능력의 불변성 개념은 피험자는 고유한 능력수준이 있기 때문에 다른 검사도구를 택하더라도 피험자의 능력은 동일하게 추정된다는 개념이다.

문항 반응이론에는 두가지 기본가정으로 단일차원성(unidimensionality)과 지역독립(local independence)이 있다. 단일 차원성의 가정이란 한 검사에 속해 있는 모든 문항들은 단일한 특성을 측정한다는 것이다. 일차원성의 가정은 검사문항의 요인분석(factor analysis)을 통하여 적합도를 검증할 수 있다. 지역독립성의 가정이란 어떤 능력을 가진 피험자가 어떤 문항에 대한 응답은 다른 문항의 응답에 전혀 영향을 주지 않는다는 가정이다. 지역독립성의 가정은 일차원성과 동형적 가정이므로 역시 요인분석으로 그 적절성을 점검할 수 있다.

2. 문항 특성 곡선

문항 특성 곡선이란 피험자의 능력  $\theta$  와 문항의 답을 맞힐 확률  $p(\theta)$  와의 함수관계를 표현한 곡선이며 <그림 3-1>과 같이 S자형을 이룬다. 각각의 문항마다 고유한 특성을 지닌 각기 다른 형태의 문항 특성 곡선을 그린다.



<그림 3-1> 문항특성곡선

문항 고유의 특성이란 문항 난이도(b)와 문항 변별도(a) 문항 추측도(c)를 들 수 있다.

문항 난이도란 능력최도에 따라 곡선의 위치를 결정하는 위치모수이며, 문항 곤란도라고 한다. 문항 난이도 모수의 범위는 일반적으로 -2에서부터 +2 정도 까지 존재하고, 문항의 답을 맞힐 확률이 0.5에 해당하는 능력 수준의 점을 문항 난이도라고 한다 <그림 3-1 참조>. 어려운 문항 일수록 곡선의 우측에 위치하게 된다.

절대적인 것은 아니지만 문항 난이도 해석을 언어적으로 표현해보면 다음과 같다.

<표 3-3> 문항난이도의 언어적 해석

문항 난이도 지수	언어적 표현
-2.0 이하	매우 쉽다
-2.0 - -.5	쉽다
-.5 - +.5	중간이다
+.5 - +2.0	어렵다
+2.0 이상	매우 어렵다

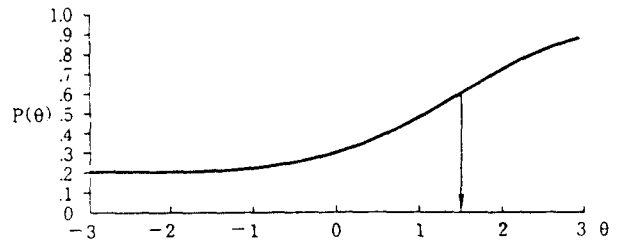
문항 변별도는 문항 특성 곡선상의 최고지수이다. 최고지수란 위치지수 아래에 있는 피험자와 위치지수 위에 있는 피험자를 변별하는 정도를 나타내며 문항 난이도를 나타내는 점에서의 곡선의 기울기에 비례한다<그림 3-1 참조>. 일반적으로 문항 변별도 모수는 0에서 부터 +2까지 존재한다.

문항 변별도의 언어적 표현에 대응하는 문항 변별도 지수의 범위를 엄밀하게 설정할 수 없으나 Baker (1985)는 다음과 같이 설정하였다.

<표 3-4> 문항 변별도의 언어적 해석

언어적 표현	정규오차이브모형	로지스틱모형
없다	.00	.00
거의 없다	.01 - .19	.01 - .34
낮다	.20 - .38	.35 - .64
적절하다	.39 - .79	.65 - 1.34
높다	.80 - .99	1.35 - 1.69
매우 높다	1.0 이상	1.70 이상
완벽하다	$+\infty$	$+\infty$

문항추측도는 문항 특성 곡선의 최저 한계이며 우연에 의하여 문항의 답을 맞힐 확률을 말한다. 문항 추측도 c값이 높을수록 그 문항은 좋지 않은 문항이며 문항을 맞힐 수 있는 확률의 범위는 c에서 1.0 이므로, 문항추측도가 높을수록 문항변별력은 떨어진다. 문항추측도의 이론적 범위는 0.0에서 1.0까지 이나 일반적으로 1/보기수 보다 낮은 확률을 나타낸다.<그림 3-2>



<그림 3-2> b=1.5, a=1.3, c= .2인 문항의 문항특성곡선

### 3. 문항 반응 모형

문항 반응 모형을 함수의 관계로 설명하기 위하여 크게 두가지 모형을 들 수 있다. 하나는 정규 오자이브 모형(Normal ogive model)이며 다른 하나는 로지스틱 모형(Logistic model)이다. 문항의 모수를 고려한 문항 반응 모형은 문항 난이도만 고려한 1-모수 모형(Rasch 모형), 문항 난이도와 문항 변별도를 고려한 2-모수 모형, 문항 난이도와 문항 변별도 그리고 문항 추측도까지 고려한 3-모수 모형이 있다.

2-모수 정규 오자이브 모형과 3-모수 정규 오자이브 모형, 2-모수 로지스틱 모형과 3-모수 로지스틱 모형은 차례로 다음과 같다.

$$P(\theta) = \int_{-\infty}^{a'(\theta-b)} \frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2} dz$$

$$P(\theta) = c + (1-c) \int_{-a}^{a'(\theta-b)} \frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2} dz$$

- e; 피험자 능력수준
- a; 문항변별도 모수
- b; 문항난이도 모수
- c; 문항추측도 모수

오자이브 모형에 의한 문항변별도는  $a^*$ 로 표기하고 로지스틱 모형에 의한 문항변별도는  $a$ 로 표기한다. 이 때,  $a = 1.702a^*$ 의 관계가 성립한다.

$$P(\theta) = \frac{1}{1 + \exp^{-L}} = \frac{1}{1 + \exp^{-a(\theta-b)}}$$

$$P(\theta) = c + (1-c) \frac{1}{1 + \exp^{-L}} \\ = c + (1-c) \frac{1}{1 + \exp^{-a(\theta-b)}}$$

위의 식에서 추측모수  $c$ 가 0이 아니라고 하면  $b$ 는 정답을 한 확률이  $c$ 와 1.0 중간인 능력수준과 같고 모수  $a$ 는 변곡점에서 곡선의 기울기에 비례한다. 기울기는 실제로  $.25a(1-c)$ 이다.

### 4. 문항모수치 추정

관찰 가능한 피험자의 검사결과를 통해 검사

의 잠재된 특성을 추정하려면, 검사를 구성하는 문항특성 추정이 선행되어야 한다. 이를 문항모수치 추정이라고 하는데 그 절차를 살펴보자.

우선 피험자들의 능력수준을 알고 있다는 가정하에 피험자 능력 집단에서 정답반응 비율을 관찰한 후 그것을 가장 적절하게 대표하는 문항특성곡선을 그린다. 문항특성곡선을 찾기 위하여 사용되는 수학적 방법에는 가장 보편적인 최대우도 추정법(maximum likelihood estimation)에 근거한다. 이 방법에 의하면 문항모수치들이 추정될 때 문항모수에 초기값들(Initial values)이 선형적으로(priori) 부여된다.

그 다음에 문항모수의 추정치를 가지고 문항 반응 모형 공식에 의하여 각 능력 수준 별로 문항의 답을 맞힐 확률  $p(\theta)$ 을 계산한다. 관찰된  $p(\theta)$  값과 문항반응 모형에 의하여 이론적으로 계산된  $p(\theta)$  값의 일치 정도는 모든 능력수준 집단에서 고려되며 그 불일치 정도가 최소화 될 때까지 모수 추정치들을 변화 시키는 과정이 계속된다. 문항모수의 추정치가 정해지면 각 능력수준에서 문항의 답을 맞힐 확률  $p(\theta)$ 가 계산되며, 이에 따라 문항특성 곡선을 그릴 수 있다.

고려할 점은 어떤 문항특성곡선이 문항에 대한 실제적인 문항 반응자료에 적합한가 하는 점이다. 그 일치 정도는 chi-square 적합도 지수(goodness-of-fit index)에 의해 추정되며, 얻어진 적합도지수의 값이 준거값(criterion value)보다 크면 문항특성곡선은 자료와 맞지 않는다고 분석한다.

문항모수를 추정하는 컴퓨터 프로그램은 BICAL, LOGIST, BILOG가 있는데 BICAL은 1-모수 Logistic 모형의 무조건적 최대우도 추정방법(UMI)프로그램이고 LOGIST는 1-모수 2-모수 3-모수 로지스틱 모형의 결합 최대우도 추정방법 프로그램이며 BILOG는 결합 최대우도 추정방법의 단점을 보완한 주변 최대우도 추정법(MML)에 의해 문항모수를 추정한다. BILOG 프로그램은 개인용 컴퓨터를 사용할 수



있다는 편리함 때문에 최근에 많이 이용되고 있다.

5. 능력모수치 추정

관찰 가능한 검사 결과를 통해 피험자의 잠재된 능력을 추정하는 것을 능력모수치 추정이라고 하며 최대우도추정법을 사용한다. 그 절차를 살펴보면,

우선 문항의 모수치를 알고있다는 가정하에 피험자 능력의 초기값을 설정하여 피험자가 각 문항의 답을 맞힐 확률을 계산한다.

그 다음 피험자의 응답에 따라 0점 혹은 1점을 부여한 문항반응 벡터와 위에서 계산한 확률의 차이를 모든 문항에 대해서 구한 뒤, 그 합이 극소화 될 때까지 전단계에서 사용한 피험자의 능력의 추정치를 교정한다.

피험자의 능력을 추정하는 공식은 다음과 같다.

$$\theta_s = \theta_s + \frac{\sum_{i=1}^N a_i [U_i - P_i(\theta_s)]}{\sum a_i^2 P_i(\theta_s) Q_i(\theta_s)}$$

$\theta_s$ ; s번째 반복교정을 통해 얻은 능력추정치

$a_i$ ; 문항 i의 변별도

$U_i$ ; 피험자의 문항 i에 대한 응답

$U_i = 1$  문항 i를 맞힌 응답

$U_i = 0$  문항 i를 틀린 응답

$P_i(\theta_s)$ ; s번째 반복해서 추정된 능력

$\theta_s$  값을 가진 피험자가 문항 i를 맞힐

확률

$Q_i(\theta_s); 1 - P_i(\theta_s)$

6. 정보함수

문항반응 이론에서 검사를 실시하는 근본적인 목적은 피험자가 소유하고 있는 능력이 능력척도상의 어디에 위치하는 가를 밝히는 것이다. 이것을 피험자능력 추정이라고 한다. 피험자의 능력이 얼마나 정확하게 추정되었는가 그 정밀성의 정도를 말해주는 것은 능력추정의 표

준오차이다. 능력추정의 표준오차란 피험자의 진짜능력(true value)을 중심으로 능력추정치들의 흠어져 있는 정도이다. 표준오차가 크면 흠어진 정도가 크다는 것이므로 피험자능력이 정확하게 추정되지 않았다는 것을 의미한다.

능력추정의 표준오차를 Fisher가 정의한 정보와 연관시켜보자. Fisher는 정보란 모수치를 추정할 때 정확성의 역이라 하였다. 이에 근거하여 정보의 개념은 능력모수치에 대한 추정치의 분산의 역으로 정의될 수 있다. 즉, 모수치를 정확하게 추정했다면 능력추정치의 표준오차가 작아지고 모수치에 관한 정보를 더 많이 갖게 된다.

정보의 양을 능력에 따라 그리면 정보함수 그래프가 되는데 능력수준이 얼마나 정확하게 추정되고 있는가를 제시한다. 문항반응 이론에서의 정보함수는 각 능력 수준에서의 능력추정치에 대한 정확성을 제시하므로 고전검사 이론에서의 신뢰도와 대등한 개념이다.

하나의 문항에 근거를 두고 피험자의 능력에 따라 정보의 양을 계산한 것이 문항정보함수이다. 각 문항반응 모형하에서의 문항정보 함수들은 다음<표 3-5>와 같다.

< 표 3-5 > 문항 정보 함수

	정규오자이브 모형	로지스틱 모형
1-모수	$\frac{h_i^2}{P_i Q_i}$	$P_i Q_i$
2-모수	$\frac{(a_i^* h_i)^2}{P_i Q_i}$	$a_i^2 P_i Q_i$
3-모수	$\frac{[a_i^* (1-c_i) h_i^2]}{P_i Q_i}$	$a_i^2 \frac{Q_i [P_i - c_i]}{P_i [1 - c_i]}$

$a_i^* = 1.702 a_i$ ; 정규 오자이브모형에 의한 문항 변별도

$a_i$ ; 로지스틱 모형에 의한 문항 변별도

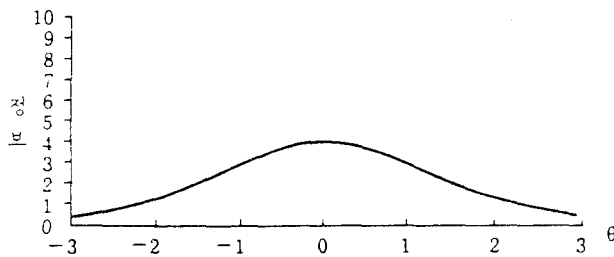
<표 3-5>에 의하면 문항변별도가 높을수록 문항정보량은 많아지고 능력수준이 문항난이도

와 일치할 때 문항정보량은 최대가 된다. 즉, 피험자의 능력과 문항난이도가 일치하면 그 문항은 피험자의 능력을 가장 정확하게 추정한다.

모든 문항에 근거를 두고 피험자의 능력에 따라 정보의 양을 계산한 것이 검사정보함수이다. 문항 하나로 피험자의 능력을 추정하는 일은 드물기 때문에 피험자능력 추정에 대한 정보에 있어서 활용상의 관심은 검사정보함수에 있다. 주어진 능력수준에서의 검사정보는 그 능력수준에서의 문항정보들의 총합이다. 그러므로 다음과 같이 정의한다.

$$I(\theta) = \sum_{i=1}^n I_i(\theta)$$

검사에 대한 정보의 양은  $\theta$ 에 대한 함수이며, 검사의 문항모수치들의 합성된 값에 의존한다. 즉, 검사정보곡선의 모양은 능력척도상의 문항난이도 분포와 검사문항의 변별도 분포의 평균값에 의하여 변한다. 검사의 문항난이도들이 밀집하여 있을 때는 검사정보곡선이 그 지점에서 높게 나타나지만 검사의 문항난이도들이 넓게 분포될 때 검사정보 곡선은 평평해지는 경향이 있다. 능력수준에 따른 검사정보를 나타내는 검사정보곡선은 <그림 3-3>과 같다.



<그림 3-3> 검사정보함수

검사정보함수의 특징을 살펴보면 첫째, 문항의 수가 많을수록 검사정보의 양은 많다.

둘째, 문항변별도가 높을수록 검사의 문항변별도 평균이 높고 검사정보량이 많다.

셋째, 검사난이도에 해당하는 능력수준에서 검사정보량은 최대이다. 또한 검사의 문항난이도들이 밀집되어 있을 때 정보량은 절정을 이룬다.

네째, 추측요인이 작용할 때 정보량은 적어진다.

다섯째, 능력추정치와 표준오차가 적을수록 정보량도 커진다.

## IV. 연구 방법 및 절차

### A. 표본 선정 및 배정

서울시 학군도를 바탕으로 학군이 서로 다른 16개 중학교를 추출한 후 93학년도 1학년 1, 2학기 중간고사, 기말고사 총 4번의 시험범위를 살펴보았다.

그 중에서 1학기 기말고사의 시험범위가 일치하는 세 학교를 택하여 학교1(동작), 학교2(동대문), 학교3(강남)이라 정하였다.

현재 중학교 수학은 5종 교과서를 택하고 있으나 추출한 16개 학교는 대체로 세 종류의 교과서를 쓰고 있었다. 표본으로 선정된 세 학교도 각각 지학사, 교학사, 동아출판사로 종류는 다르지만 단위 구성 상 큰 차이를 보이고 있지 않아 선택의 어려움은 없었다.

그러나 시험 범위는 학군간, 학교간에 차이를 보였다. 학군간의 차이가 나는 이유는 진도 문제이다. 강남지역의 학교는 진도가 빠르기 때문에 일차방정식-함수의 그래프까지가 시험범위인 반면 북부, 동부, 동작, 서부지역 학교의 시험 범위는 유리수-일차방정식의 활용까지였다.

학교간의 차이가 나는 이유는 학교 운영상의 문제이다. 학교의 각종 행사로 인해 진도의 늦고 빠름의 차이가 나고, 또 학교에 따라서는 1년의 4번 정기 시험이외에 학교 자치내에서 실력 향상을 위해 영 수 과목을 중심으로 2-3번의 시험을 더 치루고 있기 때문에 시험범위의

차이가 난다.

그 외에 학생들이 응용 활용 면에 약함을 감안하여 유리수-일차방정식 활용 전 또는 일차방정식-함수의 그래프 전 까지 소단원을 나누어 시험범위로 정하는 학교도 있었다. 그러나 이것은 총합 평가의 성격을 띤 기말고사로서 합당한 지 의문이었다.

B. 문항수와 피험자수

학교에 따른 문항수와 피험자수는 다음과 같다.

< 표 4-1 > 문항수와 피험자수

학 교	문항수		피험자수		
	객관식: 4지선다	주관식: 단답형	남	여	계
학교1(동작)	19	6	49	50	99
학교2(동대문)	18	7	108		108
학교3(강남)	15	10	50	48	98
전 체	52	23	207	98	305

각 학교에 따라서, 93년 1학년 1학기 기말고사 성적이 최상위 혹은 최하위 학급을 제외하 나머지 학급에서 2학급을 선택하여 모두 6학급을 선정하였다.

C. 분석방법

BILOG 프로그램을 통한 문항 모수치 추정방법을 택하여 문항 분석을 하였다.

PC-BILOG 프로그램은 크게 세부분으로 나누어져 있는데, 첫째 부분은 컴퓨터 언어로 전환하여 둘째, 셋째 단계를 수행할 수 있도록 임시화일(temporary file)을 만들고 고전검사 이론에 의한 문항난이도(정답비율)와 변별도 지수(피어슨 상관계수와 문항 검사 양분점 상관계수)를 산출한다. 둘째 부분은 문항반응 이론에 의한 문항특성 추정 방법의 하나인 주변 최대

우도추정법(Marginal Maximum Likelihood Estimation)으로 문항 모수치를 추정한다. 셋째 부분은 피험자 능력을 추정하고 능력추정의 표준오차를 제시한다. 검사정보 함수와 그래프, 그리고 선택사항에 따라 신뢰도 계수도 제시한다.

문항의 모수치를 추정할 때, 정규 오자이브 모형 혹은 로지스틱 모형을 선택할 수 있으며 그에 따른 1-모수 2-모수 그리고 3-모수 모형에 의하여 문항의 모수치를 추정할 수 있다. 본 연구에서는 3-모수 로지스틱 모형을 선택하여 문항모수치와 능력모수치를 추정하였다.

V 분석결과

A. 고전검사 이론에 의한 문항분석

1. 피험자 점수의 서술통계

93년 1학기 기말고사 점수를 학교별로 제시하면 다음 <표 5-1>과 같다. 평균, 표준편차, 범위는 문항배점 4점씩에 의한 것이고, Cronbach  $\alpha$  신뢰도 계수와 주관식 객관식 상관계수는 문항배점 1점씩에 의한 것이다.

<표 5-1> 문항배점에 의한 학교별 평균, 표준편차, 범위, 신뢰도 계수, 상관계수

학 교	평균	표준 편차	최고 점수	최저 점수	신뢰 도	상관 계수
학교1(동작)	62.545	22.089	100	20	.884	.767
학교2(동대문)	54.815	25.393	100	12	.911	.853
학교3(강남)	51.837	28.284	100	0	.925	.751

본 연구에서는 주, 객관식 상관계수를 산출하기 위하여 다음 식을 사용한다.

$$r_{XY} = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2] - [N \sum Y^2 - (\sum Y)^2]}}$$

X; 객관식 점수

Y; 주관식 점수

N; 피험자수

세 학교 모두 문항수는 25문항이며, 주관식 문항수는 학교1이 6문항 학교2가 7문항 학교3이 10문항이다.

평균을 살펴보면, 학교1이 가장 높고 학교3, 학교2의 순으로 나타났다. 특히 학교1과 학교3은 10.708로 큰 차이를 보였고 학교1과 학교2도 7.73의 차이를 보였다.

표준편차는 학교1이 가장 적고 학교2, 학교3의 순으로 나타났다.

신뢰도는 세 학교 모두 .85이상으로 측정하려는 것을 상당히 정확하게 오차없이 측정하고 있음을 나타냈고 세 학교 중 학교3이 가장 신뢰도가 높은 것으로 나타났다.

주관식 객관식 상관계수를 살펴보면 학교2가 가장 높고 학교1, 학교3은 비슷한 수준이었다. 그 차이는 학교1과 학교2가 .086으로 학교2가 높았고, 학교2와 학교3은 .102로 학교2가 높았으며, 학교1과 학교3은 .016으로 학교1이 높았다.

2. 답지반응분포에 따른 오답지 매력도

93년도 1학년 1학기 기말고사의 답지반응분포를 살펴본 후, 부적절한 문항으로 조사된 것을 학교별로 제시하면 다음 <표 5-2>와 같다.

<표 5-2> 답지반응분포에 따른 부적절한 문항

학교	문항번호	정답	답지반응률(%)				
			1	2	3	4	무반응
학교1 (동작)	13	4	48.48	16.16	6.06	26.26	
학교2 (동대문)	5	2	30.56	35.19	12.04	15.74	6.48
학교3 (강남)	9	2	30.61	21.43	37.76	8.16	2.04

학교1의 문항13은 정답4의 반응 비율보다 오답1의 반응비율이 22.22% 더 높게 나타났다. 문항이 어려울 때 이런 결과가 흔히 나타나는데 문항13은 수정을 요한다. 학교3의 문항9도 이와 비슷한 경우이다. 학교2의 문항5는 무

반응이 6.48%로 피험자 108명 중 7명이 반응하지 않은 문항이다. 위 경우와 마찬가지로 어려운 문항으로써 수정을 요하는 문항이다.

이와 같이 각 학교의 기말고사는 1개씩의 수정을 요하는 문제가 있었다.

93년도 1학년 1학기 기말고사의 문항별 오답지 매력도의 결과를 학교별로 제시하면 <표 5-3>과 같다. 문항별 오답지 매력도는 객관식 문항에 해당하므로 여기에서는 객관식 문항의 결과를 제시하였다.

본 연구에서 각 문항의 오답지 반응분포의 양호도를 분석하기 위해 사용한 기준은 3개의 오답지 매력도 지수가 모두 양호할 경우는 양호한 문항, 3개의 오답지 중 2개의 오답지 매력도 지수가 양호할 경우에는 판단유보, 그리고 오답지 매력도 지수가 양호한 오답지 수가 1개 이하인 경우에는 불량한 문항으로 해석을 하였다.

<표 5-3> 문항별 오답지 매력도

학교	객관식			문항수
	양호 (%)	판단유보 (%)	불량 (%)	
학교1 (동작)	12 (63.16)	6 (31.58)	1 (5.26)	19
학교2 (동대문)	10 (55.56)	7 (38.89)	1 (5.56)	18
학교3 (강남)	9 (60)	6 (40)	0 (0)	15

문항별 오답지 매력도가 양호한 문항은 학교1이 제일 많고 학교3이 제일 적었으며, 불량한 문항은 학교1과 학교2가 각각 1개씩을 가지고 있었다. 판단유보의 문항은 세 학교가 비슷한 수준이었다.

3. 문항난이도와 문항변별도 측정치에 대한 서술 통계

고전검사 이론에 의한 문항분석을 위하여 문항난이도는 정답비율, 문항변별도는 양분점 상관계수로 측정하였다.

93년 1학기 기말고사 점수의 문항난이도와 문항변별도 추정치의 평균, 표준편차, 범위는 <표 5-4>와 같다.

<표 5-4> 문항난이도와 문항변별도 추정치에 대한 평균, 표준편차, 범위

학 교	문항난이도				문항변별도			
	평균	표준편차	최고치	최저치	평균	표준편차	최고치	최저치
학교1 (동작)	.614	.189	.859	.212	.585	.126	.832	.355
학교2 (동대문)	.548	.164	.815	.213	.637	.154	.916	.348
학교3 (강남)	.518	.155	.765	.163	.698	.112	.874	.503

문항난이도를 살펴보면, 학교1이 가장 어렵고 학교2, 학교3의 순으로 나타났으나 세 학교 모두 평균 난이도 .5정도의 적당한 수준을 나타냈다.

문항변별도를 살펴보면, 학교3이 가장 높고 학교2, 학교1의 순으로 나타났다. 상관계수의 언어적 해석에 따르면 .90에서 1.00은 아주 상관성이 높다, .70에서 .90은 상관성이 높다, .40에서 .70은 확실히 상관성이 있다, .20에서 .40은 상관성이 있기는 하나 낮다, .00에서 .19는 상관성이 거의 없다이므로 세 학교 모두 확실히 상관성이 있는 것으로 나타났다.

B. 문항반응 이론에 의한 문항분석

1. 문항모수 추정치에 대한 서술통계

93년 1학기 기말고사 점수의 모든 문항에 대한 문항난이도 모수, 문항변별도 모수, 문항추측도 모수는 PC-BILOG 프로그램으로 추정하였다. <표 5-5>는 문항모수치의 평균, 표준편차, 범위이다.

전체 문항난이도 추정치들의 평균을 비교해보면, 학교2(동대문)가 .371로 가장 어렵었고 그 다음이 학교3(강남), 학교2(동작)의 순으로 나타났다.

<표 5-5> 문항모수 추정치의 평균, 표준편차, 범위

학 교	문항난이도			
	평균	표준편차	최고치	최저치
학교1 (동작)	-.298	.790	1.386	-1.482
학교2 (동대문)	.371	.695	1.466	-.919
학교3 (강남)	.065	.628	1.634	-.883

학 교	문항변별도			
	평균	표준편차	최고치	최저치
학교1 (동작)	1.031	.316	1.760	.573
학교2 (동대문)	1.422	.433	2.185	.706
학교3 (강남)	1.225	.334	2.131	.735

학 교	문항추측도			
	평균	표준편차	최고치	최저치
학교1 (동작)	.148	.035	.198	.050
학교2 (동대문)	.171	.071	.300	.049
학교3 (강남)	.062	.020	.114	.030

절대적 기준은 아니나 문항난이도의 언어적 표현에 의하면 매우 쉬운 문항은 문항난이도가 -2.0이하, 쉬운 문항은 -2.0에서 -0.5, 중간정도의 문항은 -0.5에서 0.5, 어려운 문항은 +0.5에서 +2.0, 매우 어려운 문항은 +2.0이상이기 때문에 세 학교 모두 중간정도의 시험이었다고 할 수 있다.

전체문항 변별도 추정치들의 평균을 비교해보면, 학교2(동대문)가 가장 높고 학교3(강남), 학교1(동작)의 순으로 나타났다.

로지스틱 모형에서의 문항변별도 평가기준에 의하면 변별력이 없으면 0, 거의 없으면 0.01에서 0.34, 낮으면 0.35에서 0.64, 적절하면 0.65에서 1.34, 높으면 1.35에서 1.69, 매우 높으면 1.70이상이기 때문에 학교1(동작), 학교3(강남)의 시험은 변별력이 적절하고, 학교2(동대문)의

시험은 변별력이 높은 것으로 나타났다.

문항추측도의 이론적 범위는 0.0에서 1.0 까지 이나 일반적으로 1/보기수 보다 낮은 확률을 나타낸다는 것을 고려할 때 전체평균 문항추측도는 세 학교 모두 낮은 편이다. 그러나 세 학교의 문항추측도 최고치를 살펴보면, 학교2(동대문)에 0.25 이상의 추측도를 가진 문항이 포함되어 있음을 알 수 있다.

2. 고전검사 이론에 의한 문항난이도, 변별도 측정치와 문항반응 이론에 의한 문항난이도, 변별도 추정치와의 비교

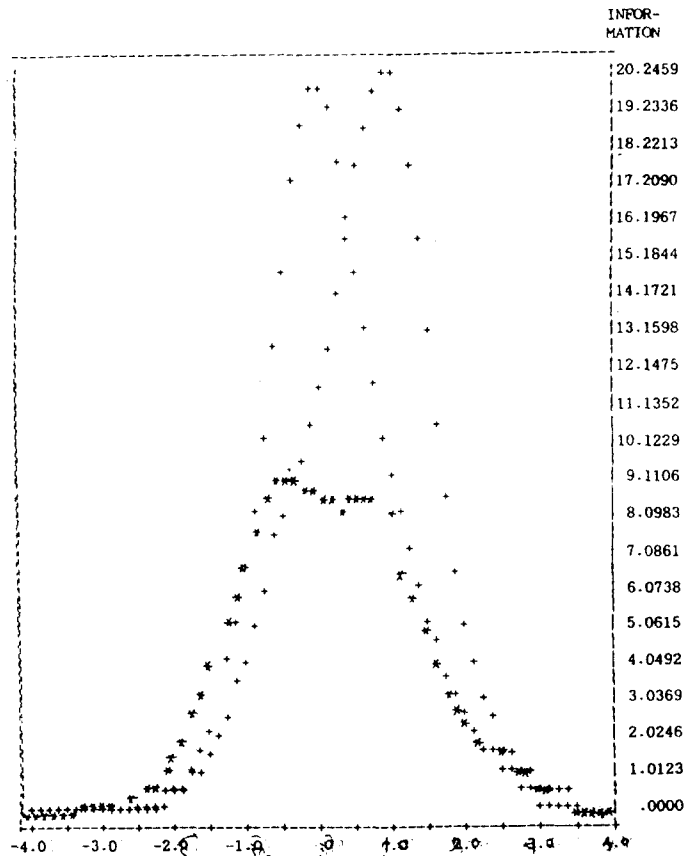
두 이론에 의한 문항난이도 문항변별도 평균치를 비교하면 다음 <표 5-6>과 같다.

<표 5-6> 두 이론에 의한 문항난이도 문항 변별도 평균치 비교

학 교	문항난이도		문항변별도	
	고전 검사	문항 반응	고전 검사	문항 반응
학교1(동작)	.614	-.298	.585	1.031
학교2(동대문)	.548	.371	.637	1.422
학교3(강남)	.518	.065	.698	1.225

문항난이도는 고전검사 이론에 의하면 학교1이 제일 높고 학교2 학교3의 순으로 나타났으나 문항반응 이론에 의하면 학교2가 제일 높고 학교3, 학교1의 순으로 나타났다.

문항변별도는 고전검사 이론에 의하면 학교3



이 제일 높고 학교2, 학교1의 순으로 나타났으나 문항반응 이론에 의하면 학교2가 제일 높고 학교3, 학교1의 순으로 나타났다.

3. 피험자 능력추정의 정확성

93년 1학기 기말고사 피험자들의 능력추정의 정확성을 학교별로 비교하기 위해 능력모수 추정치와 검사정보함수를 PC-BILOG프로그램으로 추정하였다. 능력모수의 평균, 표준편차, 범위는 <표 5-7>과 같다.

<표 5-7>능력모수 추정치의 평균, 표준편차, 범위

학 교	능력모수추정치			
	평균	표준편차	최고치	최저치
학교1(동작)	-.298	.790	1.386	-1.482
학교2(동대문)	.371	.695	1.466	-.919
학교3(강남)	.065	.628	1.634	-.883

학교1의 피험자수는 99, 학교2의 피험자수는 108, 학교3의 피험자수는 98이며, 피험자들의 능력은 대체로 비슷한 수준으로 추정되었다.

피험자 능력추정의 정확성을 재기위하여 능력수준 -2.0, -1.5, -1.0, -.5, 0, .5, 1.0, 1.5, 2.0에서의 검사정보함수를 산출하였다. 그 결과는 <표 5-8>과 같고 각 능력수준에서의 검사정보함수를 도식화 하면 <그림 5-1>과 같다.

<표 5-8> 검사정보 함수

능력수준	학교1(동작)	학교2(동대문)	학교3(강남)
-2.0	1.8	0.5	1.0
-1.5	4.5	2.5	2.5
-1.0	6.8	6.1	7.0
-.5	9.0	9.3	12.8
0	8.8	12.9	19.6
.5	8.6	17.9	14.8
1.0	8.1	20.2	9.7
1.5	4.9	8.6	5.5
2.0	2.3	2.7	2.9

학교1과 학교2를 비교하면 능력수준 -2.0에서 -1.0까지는 학교1이 학교2보다 검사정보가 많았고 능력수준 -.5에서 2.0까지는 학교2가 학교1보다 검사정보가 훨씬 많았다. 최대검사 정보는 학교1이 능력수준 -.5에서 9.0이고 학교2가 능력수준 1.0에서 19.3으로 10.3의 큰 차이를 보였다.

학교2와 학교3을 비교하면 능력수준 -2.0에서 .0까지는 학교3이 학교2보다 검사정보가 많았고 능력수준 .5에서 2.0까지는 학교2가 학교3보다 검사정보가 많았다. 최대검사 정보는 학교2가 능력수준 1.0에서 20.2이고 학교3이 능력수준 .0에서 19.6으로 그 차이는 .6이었다.

학교1과 학교3을 비교하면 능력수준 -2.0과 -1.5를 제외한 모든 능력수준에서 학교3이 학교1보다 검사정보가 훨씬 많았다. 최대검사 정보는 학교1이 -.5에서 9.0이고 학교3이 .0에서 19.6으로 10.6의 큰 차이를 보였다.

VI. 결 론

본 연구의 결론은 다음과 같다.

첫째, 고전검사 이론에 의해 문항분석을 한 결과, 세 학교 검사지의 난이도는 .5정도의 적당한 수준을 나타냈고 변별도는 세 학교 모두 확실히 상관이 있는 것으로 나타났다.

둘째, 각 학교의 평균은 학교1(동작)이 가장 높고 학교2(동대문) 학교3(강남)의 순으로 나타났고 이에 반해 각 학교의 신뢰도는 학교3(강남)이 가장 높고 학교2(동대문) 학교1(동작)의 순으로 나타났다. 평균이 가장 낮게 나타난 학교3(강남)이 제일 높게 나타난 학교1(동작)보다 피험자의 능력을 더 정확하게 측정하고 있음을 나타냈다.

셋째, 문항반응 이론에 의해 문항분석을 한 결과, 세 학교 검사지의 난이도는 모두 중간정도였고 변별도는 학교1(동작) 학교3(강남)은 적절하고 학교2(동대문)는 높은 것으로 나타났다. 추측도는 세 학교 모두 낮았으나 학교2(동대문)

의 경우 .25 이상의 추측도를 갖는 문항이 있었다.

네째, 검사정보함수에 의하면 학교2(동대문)와 학교3(강남)은 거의 모든 능력수준에서 학교1(동작)보다 검사정보가 많았다. 특히 최대검사정보는 2배 이상의 큰 차이를 보였다. 이것은 학교1에 비해 학교2 학교3이 피험자의 능력을 훨씬 더 정확하게 추정하고 있음을 나타낸다. 학교2는 능력수준 1.0에서 최대검사정보를 보이고 학교3은 능력수준 .0에서 최대검사정보를 보이므로 피험자들의 능력을 가장 정확하게 추정하는 능력수준은 학교2가 학교3에 비해 더 높은 것으로 나타났다.

이상의 결과는, 각 학교별로 평균에 의해 석차를 정하고 석차에 따라 내신등급을 정하는 일률적인 내신등급 반영이 합당치 못함을 보여준다. 따라서 학업성취의 절대적 기준과 그것을 측정할 수 있는 공신력 있는 제도가 마련되어야 한다.

## 참 고 문 헌

- 국립교육평가원 (1993). 교육평가 제 3호.  
 배호순 (1990). 평가의 원리. 교육과학사.  
 박윤근, 양진숙, 임형 (1993). 교육과정 인정 시험결과 분석연구. 국립교육평가원.  
 성태제 (1991). 문항반응이론 입문. 양서원.  
 오성삼, 김성훈, 성태제, 김경석 (1992). 문항제작과 문항분석방법. 교육평가연구회.  
 황정규 (1993). 학교학습과 교육평가. 교육과학사.  
 추정아 (1992). 대학수학능력시험 실험평가의 성별에 따른 차별기능문항 추출.  
 황소림 (1993). 대학수학능력시험 제 6차·7차 실험평가의 문항 특성과 피험자능력 동등화.  
 정범모 (1993). 입시와 교육개혁. 나남출판사.  
 송미영 (1993). 이분반응모형과 등급반응모형에 의한 문항특성과 피험자능력 추정치의 정확성.