

## 모의실험에 의한 가교피험자설계의 조건에 따른 동등화 방법 비교<sup>1)</sup>

안 수 현(국립국제교육원 연구원)\*

김 현 철(성균관대학교 교수)\*\*

### <요 약>

본 연구는 모의실험을 통해 가교피험자설계에서 여러 동등화 방법들의 동등화 결과를 비교하는 것을 목적으로 한다. 가교검사설계 동등화 방법에 대한 선행연구와 34회·35회 한국어능력시험(TOPIK)을 분석한 자료를 토대로 검사에 포함되는 문항의 극단값 조건, 난이도 극단값을 포함하는 문항의 비율 조건, 전체 피험자 수 조건, 가교피험자 비율 조건을 설정하고, 조건별로 100회의 모의실험(simulation)이 실시되었다. 동등화 수행 과정에 포함된 동등화 방법들은 문항반응이론(IRT) 진점수 동등화, 선형 동등화, 사전완곡화(로그 선형) 동백분위 동등화, 비완곡화 동백분위 동등화, Circle-Arc 동등화 등 다섯 가지였다. 오차를 나타내는 통계량 RMSE가 실험조건을 독립변수로 하는 분산분석에 의하여 비교되었다. 분석에는 R과 SAS9.2가 사용되었다. 연구결과는 첫째, 모든 실험 조건에서 난이도 극단값이 +2일 때가 난이도 극단값이 -2일 때보다 오차가 매우 컸다. 난이도 극단값 -2가 포함된 조건에서는 난이도 극단값 포함 문항의 비율이 커질수록 오차가 감소하였으나, 문항의 난이도 극단값 +2가 포함된 조건에서는 난이도 극단값 포함 비율이 커질수록 오차가 커졌다. 난이도 극단값 포함 문항의 비율이 0%일 때 가교피험자 비율 조건에서 모두 오차가 작았고, 난이도 극단값을 포함하는 문항의 비율이 20%일 때 가교피험자 비율의 대부분 조건에서 오차가 가장 컸다. 모든 실험 조건에서 전체 피험자 수가 30,000명일 때는 10,000명일 때보다 오차가 감소하였고, 가교피험자 비율이 1%일 때 오차가 가장 컸으며, 5%, 10%, 15%, 20%일 때의 오차는 비슷하였다. 난이도 극단값이 -2일 경우에는 가교피험자 비율이 1%일 때 오차가 가장 컸으며, 난이도 극단값이 +2일 경우에는 모든 가교피험자 비율 조건에서 큰 차이를 보이지 않았다. 둘째, 모든 실험 조건에서 Circle-Arc 동등화 방법의 오차가 가장 작았고, 사전완곡화(로그 선형) 동백분위 동등화 방법의 오차가 가장 컸다.

주제어 : 모의실험, 고전검사이론 동등화, 문항반응이론 동등화, 검사 동등화, 동등화 오차

1) 본 연구는 성균관대학교 일반대학원(2016) ‘모의실험에 의한 가교피험자설계의 조건에 따른 동등화 방법 비교’ 박사학위 논문을 기반으로 수정·보완한 것임.

\* 제1저자, anss4608@moe.go.kr

\*\* 교신저자, hkim@skku.edu

## I. 연구의 필요성 및 목적

1997년 3천명도 안되던 한국어능력시험(TOPIK, Test Of Proficiency In Korean)의 응시자는 2013년 1월 이미 누적지원자가 100만 명을 돌파했고, 특히 개편 직전인 34회 시험에서는 국내·외 46개국 226개 시험장에서 총 7만 2,079명이 지원하였다. 이러한 한국어능력시험 지원자수의 급증에 따른 수요를 감당하기 위해 2013년까지 매년 4회 시행해 오던 시험 횟수를 2014년에는 5회, 2015년부터는 6회로 늘려 실시하고 있다. 2014년 7월에 시행된 35회 시험부터는 국가·사회적 측면에서 한국어능력시험에 요구되는 정책적 수요를 반영하고, 문법·이론 중심의 기존 시험방식에서 벗어나 언어사용능력 중심 평가로 개편되었다. 기존 한국어능력시험은 수준에 상관없이 일관되게 어휘 및 문법, 쓰기, 듣기, 읽기 등 4가지 영역으로 구성되었으나, 개편 체제에서는 어휘/문법을 언어 활용이라는 맥락적 상황에서 평가하기 위하여 다른 영역(특히, 읽기 영역)의 하위 내용으로 통합하고, 쓰기 영역은 상급 수준에 해당하는 ‘한국어능력시험 II’에만 포함된다. 개편 후 한국어능력시험의 체제는 우선, 기존 한국어 숙달 수준을 나타내는 1-6등급을 유지하며, 시험의 종류를 개편전의 초급, 중급, 고급 3종에서 한국어능력시험 I, 한국어능력시험 II의 2종으로 통합한다. 즉 기존 초급 시험은 ‘한국어능력시험 I’로 유지하며, 중급과 고급 시험유형을 ‘한국어능력시험 II’로 통합한다. 이에 따라 수준별 시험의 명칭이 변경되었을 뿐, 개편 전과 후의 시험에서 등급별 평가 기준은 동일하게 적용된다.

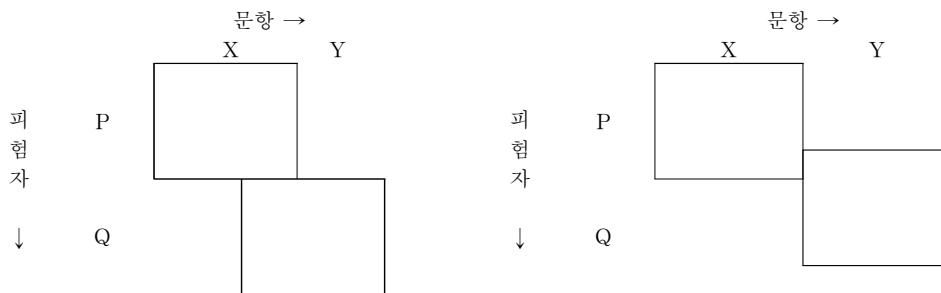
현행 한국어능력시험은 원점수와 원점수에 근거한 등급 기준으로 점수 결과가 해석되고 있다. 이러한 상황에서 시험 간 결과 해석 및 보고의 일관성을 유지하기 위해서는 무엇보다도 동일한 검사틀 내에서 난이도가 비슷하게 유지되어야 하는 것이 선결 조건이라 할 수 있으나 한국어능력시험의 현재 출제체제는 검사의 난이도를 사전에 분석 및 조정할 수 있는 사전검사(pilot test) 과정을 포함하고 있지 못하기 때문에 회차별 시험 난이도의 일관성을 유지하는 것에 근본적인 한계를 지니고 있다(민경석, 2014). 따라서 특정 시기의 시험을 치른 사람이 다른 시기의 시험을 치른 사람보다 더 유리하거나 더 불리한 상황이 발생하고 있다. 그러므로 일반적으로 한 해에 두 번 이상의 수험 기회를 부여하는 시험이나 여러 해에 걸친 학생들의 능력 수준의 변화를 알아보기 위한 대규모 시험에서는 검사 유형 간 점수의 비교 가능성을 확보하기 위해 가교 역할을 할 수 있는 문항 또는 피험자 설계가 필요한 요소이다. 그러나 한국어능력시험에는 아직 이러한 동등화 설계가 마련되어 있지 않기 때문에 향후 타당한 동등화 설계를 위한 기초자료를 제공하고자 본 연구에서는 한국어능력시험의 실제 자료 특성을 반영한 여러 가지 실험 조건에 따른 모의자료를 생성하고 이들 자료에 여러 가지 동등화 방법들을 적용한 결과를 비교하고자 한다.

## II. 이론적 배경

### 1. 동등화 설계

검사 비밀이 유지되어야 할 필요가 있거나 다른 실제적인 이유 때문에 동등화할 두 검사를 동시에 실시할 수 없을 때 비 동등집단 가교검사설계가 유용하다. 특히 두 피험자 집단이 동질적이지 못할 때 난이도를 비롯한 검사 특성의 차이와 피험자 능력의 차이가 혼동되지 않도록 하기 위하여 이 설계가 사용된다. 일반적으로 동등화할 두 검사가 동시에 실시되는 경우보다는 실시 간격이 있고, 두 피험자 집단의 동질성을 가정할 수 없는 경우가 많기 때문에 이 방법이 가장 현실적인 설계이다. 이 설계에서 가교 검사의 역할이 매우 크다. 가교 검사는 동등화 검사의 축소판이어야 한다. 즉 동등화 검사 X와 Y의 문항 모집단과 가교 검사의 문항 모집단은 같아야 한다(Klein & Jarjoura, 1985). 두 개의 동등화 검사 내에서 가교 검사 문항은 기능을 똑같이 해야 한다. 따라서 문항 내의 단어나 답지의 순서 등이 달라져서는 안되며 문항의 순서와 위치가 동일해야 한다. 집단 간 능력 차이가 지나치게 크다면 가교 검사 문항 수가 전체 문항 수의 20% 이하로서 내용 대표성을 확보하지 못할 때 동등화 오차가 커진다(Cook et al, 1988).

하지만 현실적으로 동등화할 두 검사의 가교 역할을 하는 가교문항이 여의치 않을 때에는 가교 피험자설계(Anchor Persons Design)를 통하여 동등화할 수 있다(Hambleton & Swaminathan, 1985; 남현우, 2001). 두 검사 모두를 치른 피험자 중 전집을 대표할 수 있는 피험자들의 능력 모수를 기준으로 두 검사 문항 모수치들의 재척도화를 수행할 수 있다. 가교검사설계(Anchor Item Design)와 가교피험자설계(Anchor Persons Design)를 비교하면 다음과 같다.



[그림 II-1] 가교검사설계(左)와 가교피험자설계(右)

## 2. 동등화 방법

검사점수의 동등화는 사용하는 함수에 따라 고전검사이론에 의한 동등화 방법과 문항반응 이론에 의한 동등화 방법으로 나누어지는데, 자료 수집 설계에 따라 동등화 방법의 결과가 다르게 나타난다. Lord(1975)는 비동등집단 가교검사 설계를 제외한 다른 설계에서는 고전적 동등화 방법들도 문항반응이론 동등화 방법만큼 좋은 동등화 결과를 보여준다고 하였다.

본 연구에서는 고전검사이론에 기초한 동등화의 대표적인 방법으로 선형 방법과 동백분위 방법을 사용하였고, 동백분위 방법은 이변량 점수 분포의 완곡화를 함에 있어 모형의 적용이 상대적으로 용이하다(Hanson, 1991; Holland & Thayer, 2000; Shin, 2011; Wang, 2011)고 평가를 받는 로그-선형 방법(log-linear method)을 포함하였다. 한편, von Davier, Holland, & Thayer(2004)는 동등화 함수를 직선 요소와 곡선 요소로 분해할 수 있고 곡선 요소를 위한 방법으로 Circle-Arc 방법을 사용할 수 있다고 제안하였으며, Livingston & Kim(2009)은 4지 선다형을 사용한 검사 상황에서 Circle-Arc 방법이 다른 동등화 방법보다 적절하다고 보고함에 따라 Circle-Arc 동등화 방법을 포함하였다. 문항반응이론 동등화 방법 중에서 진점수 동등화 방법은 검사가 동형검사가 아닐 때, 동등화의 기본 가정인 공정성, 집단 불변성, 대칭성의 조건들을 충족시키는 것으로 보고된(Cook & Eignor, 1983; Lord, 1980) 문항반응이론 진점수 동등화 방법이 함께 비교되었다.

## III. 연구 방법

### 1. 연구자료

본 연구에서는 모의실험을 진행하기 위한 실험의 조건들을 설정하기 위해 34회·35회 한국어능력시험 자료에 대한 분석을 진행하였다. 한국어능력시험의 전체 문항 수는 50문항이며, 총점은 100점으로 하여 모의자료를 생성하였다. 한국어능력시험처럼 다양한 능력의 피험자가 충분히 많은 경우에는 피험자의 능력이 정규분포한다(이상하, 박도영, 2012)는 가정하에 피험자의 능력수준은  $N(0, 1)$  분포를 이용하여 응답반응 자료를 생성하였다. 두 시험의 응시집단은 유사한 능력을 가진 것으로 가정하였다. 문항모수 분포는 한국어능력시험 문항모수의 실제 분포를 참조하여 변별도 모수(a)는 최솟값과 최댓값이 0.5, 2.8을 모수로 하는 균일분포( $U(0.5, 2.8)$ ), 난이도 모수(b)는 평균과 표준편차가 -0.5, 1인 정규분포( $N(-0.5, 1)$ ), 추측도 모수(c)는 최솟값과 최댓값이 0.05, 0.3을 모수로 하는 균일분포( $U(0.05, 0.3)$ )에서 추출하였다.

<표 III-1> 34회·35회 한국어능력시험 문항 난이도 및 변별도 결과\_듣기

| 문항<br>번호 | 난이도    |        |        |         |          | 변별도   |       |       |         |          |
|----------|--------|--------|--------|---------|----------|-------|-------|-------|---------|----------|
|          | 초급     | 중급     | 고급     | TOPIK I | TOPIK II | 초급    | 중급    | 고급    | TOPIK I | TOPIK II |
| 1        | -2.357 | -2.980 | -1.951 | -1.951  | -2.279   | 2.664 | 0.980 | 1.041 | 1.155   | 1.220    |
| 2        | -2.265 | -2.329 | -2.049 | -2.049  | -1.537   | 0.662 | 1.077 | 1.638 | 1.581   | 1.890    |
| 3        | -2.000 | 0.225  | 0.263  | 0.263   | -0.472   | 2.273 | 0.886 | 1.232 | 2.106   | 1.217    |
| 4        | -1.450 | -0.235 | -2.676 | -2.676  | -0.519   | 2.027 | 1.433 | 0.755 | 2.366   | 1.605    |
| 5        | -1.211 | -0.508 | -0.827 | -0.827  | 0.132    | 2.193 | 1.619 | 1.098 | 1.281   | 1.550    |
| 6        | -0.928 | -0.936 | -0.819 | -0.819  | -1.539   | 2.477 | 1.664 | 1.271 | 1.179   | 0.936    |
| 7        | -1.155 | -0.763 | -1.633 | -1.633  | -0.567   | 0.694 | 1.742 | 0.566 | 1.824   | 1.745    |
| 8        | -0.438 | -1.420 | -0.286 | -0.286  | -0.518   | 1.340 | 0.943 | 0.597 | 2.129   | 1.563    |
| 9        | -1.004 | -1.016 | -0.659 | -0.659  | -0.392   | 2.230 | 0.556 | 1.730 | 1.706   | 1.872    |
| 10       | -1.056 | -0.499 | -1.933 | -1.933  | 0.069    | 2.132 | 1.250 | 1.460 | 1.590   | 0.745    |
| 11       | -0.668 | -0.684 | 1.063  | 1.063   | -0.291   | 1.839 | 1.397 | 0.672 | 1.971   | 1.642    |
| 12       | -1.196 | -1.310 | -1.334 | -1.334  | -1.380   | 2.520 | 1.155 | 0.978 | 1.527   | 0.612    |
| 13       | -1.590 | -0.438 | -1.480 | -1.480  | -0.073   | 2.500 | 1.404 | 1.026 | 2.565   | 1.293    |
| 14       | -0.678 | -1.493 | -2.720 | -2.720  | -0.681   | 1.478 | 1.434 | 1.537 | 1.674   | 1.455    |
| 15       | -1.496 | -0.338 | -1.485 | -1.485  | 0.605    | 1.365 | 1.534 | 0.801 | 2.038   | 0.964    |
| 16       | -1.928 | -0.035 | -0.737 | -0.737  | -0.248   | 1.328 | 1.725 | 0.713 | 1.464   | 0.981    |
| 17       | -1.706 | -0.498 | -1.928 | -1.928  | -0.172   | 1.476 | 0.890 | 1.623 | 1.278   | 1.881    |
| 18       | -1.351 | -0.736 | -2.282 | -2.282  | -0.256   | 1.758 | 2.089 | 1.866 | 1.118   | 1.122    |
| 19       | -0.365 | -0.749 | -1.233 | -1.233  | -0.259   | 1.414 | 1.705 | 1.702 | 1.645   | 1.445    |
| 20       | 0.567  | -0.505 | -1.199 | -1.199  | -0.738   | 1.425 | 1.341 | 1.723 | 1.930   | 1.342    |
| 21       | 0.194  | -0.610 | 0.561  | 0.561   | -0.529   | 1.228 | 1.109 | 0.907 | 1.915   | 1.348    |
| 22       | -1.342 | -1.066 | 0.484  | 0.484   | 0.130    | 1.698 | 1.155 | 1.096 | 1.138   | 1.080    |
| 23       | -0.650 | -0.889 | -1.241 | -1.241  | -0.403   | 0.625 | 0.852 | 1.144 | 1.467   | 1.487    |
| 24       | -0.366 | -1.040 | -1.001 | -1.001  | 0.353    | 1.343 | 1.412 | 0.823 | 1.741   | 1.180    |
| 25       | 0.011  | -0.503 | -0.503 | -0.503  | 0.032    | 0.956 | 0.972 | 1.150 | 1.170   | 0.683    |
| 26       | 0.561  | 0.927  | -0.149 | -0.149  | 0.021    | 0.810 | 0.659 | 1.174 | 1.141   | 0.676    |
| 27       | 0.003  | -0.414 | 0.231  | 0.231   | 1.668    | 1.126 | 1.007 | 1.375 | 1.948   | 0.671    |
| 28       | 0.009  | -0.482 | -0.658 | -0.658  | -0.208   | 1.043 | 1.030 | 0.873 | 1.391   | 1.259    |
| 29       | -0.253 | 0.602  | -0.856 | -0.856  | -1.588   | 0.741 | 0.650 | 0.650 | 0.575   | 1.411    |
| 30       | 0.911  | -0.768 | -0.800 | -0.800  | -0.738   | 0.820 | 0.966 | 0.963 | 1.495   | 1.464    |
| 31       |        |        |        |         | -0.499   |       |       |       |         | 0.756    |
| 32       |        |        |        |         | -0.574   |       |       |       |         | 0.687    |
| 33       |        |        |        |         | -0.623   |       |       |       |         | 1.277    |
| 34       |        |        |        |         | 2.318    |       |       |       |         | 0.281    |
| 35       |        |        |        |         | 0.156    |       |       |       |         | 1.028    |
| 36       |        |        |        |         | -0.832   |       |       |       |         | 1.199    |
| 37       |        |        |        |         | 2.625    |       |       |       |         | 0.265    |
| 38       |        |        |        |         | 0.525    |       |       |       |         | 0.630    |
| 39       |        |        |        |         | 2.727    |       |       |       |         | 0.447    |
| 40       |        |        |        |         | 0.900    |       |       |       |         | 0.907    |
| 41       |        |        |        |         | 0.019    |       |       |       |         | 0.944    |
| 42       |        |        |        |         | 0.294    |       |       |       |         | 0.262    |
| 43       |        |        |        |         | 0.959    |       |       |       |         | 1.167    |
| 44       |        |        |        |         | 0.282    |       |       |       |         | 0.666    |
| 45       |        |        |        |         | -1.104   |       |       |       |         | 1.017    |
| 46       |        |        |        |         | 1.656    |       |       |       |         | 0.545    |
| 47       |        |        |        |         | -0.365   |       |       |       |         | 1.369    |
| 48       |        |        |        |         | -0.001   |       |       |       |         | 1.083    |
| 49       |        |        |        |         | 0.213    |       |       |       |         | 0.698    |
| 50       |        |        |        |         | -0.353   |       |       |       |         | 1.067    |
| 평균       | -0.873 | -0.716 | -0.995 | -0.995  | -0.021   | 1.573 | 1.221 | 1.138 | 1.603   | 1.093    |

&lt;표 III-2&gt; 34회·35회 한국어능력시험 문항 난이도 및 변별도 결과\_읽기

| 문항<br>번호 | 난이도    |        |        |         |          | 변별도   |       |       |         |          |
|----------|--------|--------|--------|---------|----------|-------|-------|-------|---------|----------|
|          | 초급     | 중급     | 고급     | TOPIK I | TOPIK II | 초급    | 중급    | 고급    | TOPIK I | TOPIK II |
| 1        | -2.620 | -1.744 | -1.911 | -3.027  | -1.928   | 1.964 | 1.190 | 1.013 | 1.176   | 1.620    |
| 2        | -1.004 | -1.722 | -1.193 | -2.439  | -0.621   | 1.850 | 2.060 | 1.047 | 2.499   | 0.839    |
| 3        | 0.894  | 1.539  | -2.053 | -3.659  | -1.275   | 0.681 | 0.617 | 1.464 | 1.036   | 2.042    |
| 4        | -1.881 | -1.306 | -1.361 | -1.714  | -0.293   | 1.557 | 1.297 | 0.588 | 1.736   | 1.857    |
| 5        | -1.595 | -0.926 | -0.675 | -2.236  | -2.025   | 2.252 | 1.089 | 1.030 | 0.953   | 1.310    |
| 6        | -1.020 | -1.682 | -0.944 | -1.619  | -2.406   | 2.151 | 1.137 | 0.932 | 2.595   | 2.166    |
| 7        | -0.983 | -1.439 | -1.445 | -1.702  | -1.308   | 1.646 | 1.163 | 1.442 | 2.251   | 2.079    |
| 8        | -1.365 | -0.694 | -1.689 | -1.092  | -0.504   | 2.641 | 1.808 | 0.653 | 1.684   | 1.645    |
| 9        | -0.410 | -1.116 | -1.225 | -1.288  | -0.671   | 1.400 | 1.047 | 1.101 | 1.474   | 1.701    |
| 10       | -0.427 | 0.252  | -1.425 | -0.987  | -0.486   | 1.493 | 1.180 | 0.932 | 1.825   | 1.403    |
| 11       | -0.766 | -0.478 | -1.167 | -1.050  | -0.504   | 1.649 | 1.031 | 0.858 | 1.707   | 1.758    |
| 12       | -0.799 | 0.169  | -0.884 | -1.143  | -0.552   | 2.047 | 0.964 | 1.301 | 1.751   | 1.501    |
| 13       | -0.763 | -0.957 | -1.180 | -0.824  | -0.653   | 1.797 | 1.478 | 1.147 | 1.364   | 1.249    |
| 14       | -0.631 | -0.958 | -1.922 | -1.178  | -0.317   | 1.409 | 1.454 | 0.636 | 1.337   | 1.495    |
| 15       | -0.717 | -1.071 | -0.015 | -0.372  | -0.530   | 2.072 | 1.244 | 1.243 | 2.337   | 1.721    |
| 16       | -0.792 | -0.719 | -1.118 | -1.431  | -0.584   | 1.961 | 1.526 | 1.385 | 2.800   | 1.264    |
| 17       | -0.767 | 0.042  | -0.619 | -0.542  | -0.026   | 2.496 | 1.025 | 1.054 | 1.227   | 1.295    |
| 18       | -0.762 | -0.129 | -1.070 | -0.965  | -0.916   | 1.908 | 0.919 | 1.170 | 1.810   | 1.386    |
| 19       | 0.144  | -0.408 | -1.418 | -0.295  | -0.509   | 0.522 | 1.624 | 1.169 | 0.632   | 0.910    |
| 20       | -0.396 | -0.085 | -0.781 | -0.798  | -0.588   | 1.178 | 1.190 | 0.896 | 1.385   | 2.007    |
| 21       | -0.271 | -0.152 | 0.411  | -0.329  | 0.368    | 1.152 | 1.350 | 0.779 | 1.263   | 0.883    |
| 22       | -0.348 | -0.140 | 0.437  | -0.234  | -0.215   | 1.115 | 1.443 | 0.871 | 1.351   | 1.877    |
| 23       | 0.136  | -0.045 | 0.051  | -0.250  | 0.348    | 1.300 | 0.886 | 1.214 | 1.247   | 0.909    |
| 24       | -0.791 | -0.083 | -0.999 | -0.832  | -0.832   | 1.498 | 1.718 | 1.126 | 1.163   | 2.329    |
| 25       | 1.560  | 0.862  | -0.564 | -0.129  | -1.016   | 0.551 | 0.863 | 1.150 | 1.359   | 1.894    |
| 26       | -0.283 | -0.311 | -0.290 | -0.698  | -0.373   | 1.460 | 0.994 | 1.176 | 1.776   | 1.903    |
| 27       | -1.009 | 0.000  | -0.379 | 0.238   | -0.556   | 1.510 | 1.417 | 1.500 | 1.277   | 1.522    |
| 28       | -0.360 | -0.550 | -0.875 | 0.082   | -0.736   | 1.442 | 1.472 | 0.914 | 0.801   | 0.827    |
| 29       | 0.464  | 0.002  | 0.709  | 0.767   | -0.677   | 0.679 | 0.944 | 0.653 | 0.849   | 2.155    |
| 30       | -0.119 | -0.249 | 0.381  | 1.519   | -0.694   | 1.133 | 0.985 | 0.605 | 0.724   | 1.996    |
| 31       |        |        |        | 0.308   | -0.031   |       |       |       | 1.043   | 1.771    |
| 32       |        |        |        | 0.683   | 0.039    |       |       |       | 1.056   | 1.405    |
| 33       |        |        |        | -0.498  | -0.556   |       |       |       | 1.570   | 1.527    |
| 34       |        |        |        | -0.164  | -0.062   |       |       |       | 0.912   | 1.349    |
| 35       |        |        |        | -0.172  | 0.252    |       |       |       | 1.101   | 1.275    |
| 36       |        |        |        | -0.250  | 0.299    |       |       |       | 0.955   | 1.174    |
| 37       |        |        |        | -0.003  | 0.256    |       |       |       | 1.533   | 1.026    |
| 38       |        |        |        | 0.788   | 0.056    |       |       |       | 0.713   | 1.195    |
| 39       |        |        |        | 0.200   | -0.277   |       |       |       | 1.021   | 1.580    |
| 40       |        |        |        | 0.752   | 0.783    |       |       |       | 0.796   | 0.689    |
| 41       |        |        |        |         | 0.354    |       |       |       |         | 1.082    |
| 42       |        |        |        |         | 0.179    |       |       |       |         | 1.246    |
| 43       |        |        |        |         | -0.131   |       |       |       |         | 1.246    |
| 44       |        |        |        |         | 2.383    |       |       |       |         | 0.530    |
| 45       |        |        |        |         | 0.767    |       |       |       |         | 0.588    |
| 46       |        |        |        |         | 1.719    |       |       |       |         | 0.530    |
| 47       |        |        |        |         | 1.565    |       |       |       |         | 0.443    |
| 48       |        |        |        |         | 1.339    |       |       |       |         | 0.772    |
| 49       |        |        |        |         | 0.849    |       |       |       |         | 1.103    |
| 50       |        |        |        |         | 0.283    |       |       |       |         | 1.066    |
| 평균       | -0.589 | -0.470 | -0.840 | -0.665  | -0.220   | 1.550 | 1.237 | 1.035 | 1.402   | 1.379    |

## 2. 실험조건

한국어능력시험의 분석 결과를 바탕으로 난이도 극단값 조건, 난이도 극단값 문항의 비율, 전체 피험자 수, 가교피험자 비율의 4가지 조건을 설정하였다.

### 가. 난이도 극단값 조건

국가단위의 시험의 경우, 다양한 학생들의 능력을 정확하게 측정하기 위하여 다양한 난이도 수준의 문항을 포함하여 검사를 구성하는 것이 일반적이다. 한국어능력시험의 문항 출제 시, 영역별 난이도 구성 지침을 마련하고 있는데, 이는 아래와 같다.

<표 III-3> 34회·35회 한국어능력시험 난이도 구성\_TOPIK I 듣기

| 난이도 | 개편 전 TOPIK I 듣기 |       |    | 개편 후 TOPIK I 듣기 |       |      | 비율(%) |
|-----|-----------------|-------|----|-----------------|-------|------|-------|
|     | 문항수(개)          | 배점(점) |    | 문항수(개)          | 배점(점) |      |       |
| 2급  | 상               | 4     | 13 | 3               | 30    | 30%  | 30%   |
|     | 중               | 7     | 24 | 6               | 60    | 60%  |       |
|     | 하               | 4     | 13 | 6               | 40    | 40%  |       |
| 1급  | 상               | 4     | 13 | 6               | 40    | 40%  | 30%   |
|     | 중               | 7     | 24 | 6               | 30    | 30%  |       |
|     | 하               | 4     | 13 | 3               | 60    | 60%  |       |
| 총합  | 30              | 100   |    | 30              | 100   | 100% | 100%  |

<표 III-4> 34회·35회 한국어능력시험 난이도 구성\_TOPIK I 읽기

| 난이도 | 개편 전 TOPIK I 읽기 |       |    | 개편 후 TOPIK I 읽기 |       |      | 비율(%) |
|-----|-----------------|-------|----|-----------------|-------|------|-------|
|     | 문항수(개)          | 배점(점) |    | 문항수(개)          | 배점(점) |      |       |
| 2급  | 상               | 4     | 13 | 4               | 30    | 30%  | 30%   |
|     | 중               | 7     | 24 | 8               | 60    | 60%  |       |
|     | 하               | 4     | 13 | 8               | 40    | 40%  |       |
| 1급  | 상               | 4     | 13 | 8               | 40    | 40%  | 30%   |
|     | 중               | 7     | 24 | 8               | 30    | 30%  |       |
|     | 하               | 4     | 13 | 4               | 60    | 60%  |       |
| 총합  | 30              | 100   |    | 40              | 100   | 100% | 100%  |

&lt;표 III-5&gt; 34회·35회 한국어능력시험 난이도 구성\_TOPIK II 듣기, 읽기

| 난이도 | 개편 전 TOPIK II 듣기, 읽기 |       |                        | 개편 후 TOPIK II 듣기, 읽기 |       |       |
|-----|----------------------|-------|------------------------|----------------------|-------|-------|
|     | 문항 수(개)              | 배점(점) |                        | 문항 수(개)              | 배점(점) | 비율(%) |
| 6급  | 상                    | 4     | 13<br>1X4=4<br>3X3=9   | 3                    | 6     | 24%   |
|     | 중                    | 7     | 24<br>3X4=12<br>4X3=12 | 5                    | 10    |       |
|     | 하                    | 4     | 13<br>1X4=4<br>3X3=9   | 4                    | 8     |       |
| 5급  | 상                    | 4     | 13<br>1X4=4<br>3X3=9   | 4                    | 8     | 26%   |
|     | 중                    | 7     | 24<br>3X4=12<br>4X3=12 | 5                    | 10    |       |
|     | 하                    | 4     | 13<br>1X4=4<br>3X3=9   | 4                    | 8     |       |
| 4급  | 상                    | 4     | 13<br>1X4=4<br>3X3=9   | 4                    | 8     | 26%   |
|     | 중                    | 7     | 24<br>3X4=12<br>4X3=12 | 5                    | 10    |       |
|     | 하                    | 4     | 13<br>1X4=4<br>3X3=9   | 4                    | 8     |       |
| 3급  | 상                    | 4     | 13<br>1X4=4<br>3X3=9   | 4                    | 8     | 24%   |
|     | 중                    | 7     | 24<br>3X4=12<br>4X3=12 | 5                    | 10    |       |
|     | 하                    | 4     | 13<br>1X4=4<br>3X3=9   | 3                    | 6     |       |
| 총합  | 60                   | 200   |                        | 50                   | 100   | 100%  |

다음 <표 III-6>, <표 III-7>에는 34회·35회 한국어능력시험의 문항 난이도와 변별도 분석결과가 제시되었다. 우선, 난이도 분포에 의하면 이들 시험은 매우 쉬운 문항(-2.0 이하)과 매우 어려운 문항(+2.0 이상)을 일부 포함하고 있으며, 문항 변별도는 적절한 범위에 고르게 분포하고 있다.



<표 III-6> 34회·35회 한국어능력시험 문항 난이도 분포

| 구분 | 시험<br>등급 | 시험<br>과목 | 문항수<br>(%) | -2.0<br>미만 | -2.0이상~<br>-0.5미만 | -0.5이상~<br>+0.5미만 | +0.5이상~<br>+2.0미만 | +2.0<br>이상 |
|----|----------|----------|------------|------------|-------------------|-------------------|-------------------|------------|
| 기존 | 초급       | 듣기       | 30(100)    | 2(6.7)     | 17(56.7)          | 9(30.0)           | 2(6.7)            | 0(0)       |
| 기존 | 초급       | 읽기       | 30(100)    | 1(3.3)     | 16(53.3)          | 11(36.7)          | 2(6.7)            | 0(0)       |
| 기존 | 중급       | 듣기       | 30(100)    | 2(6.7)     | 17(56.7)          | 9(30.0)           | 2(6.7)            | 0(0)       |
| 기존 | 중급       | 읽기       | 30(100)    | 0(0.0)     | 13(43.3)          | 15(50.0)          | 2(6.7)            | 0(0)       |
| 기존 | 고급       | 듣기       | 30(100)    | 4(13.3)    | 19(63.3)          | 5(16.7)           | 2(6.7)            | 0(0)       |
| 기존 | 고급       | 읽기       | 30(100)    | 1(3.3)     | 21(70.0)          | 7(23.3)           | 1(3.3)            | 0(0)       |
| 개정 | TOPIK I  | 듣기       | 30(100)    | 6(20.0)    | 17(56.7)          | 6(20.0)           | 1(3.3)            | 0(0)       |
| 개정 | TOPIK I  | 읽기       | 40(100)    | 2(10.0)    | 16(40.0)          | 15(37.5)          | 5(12.5)           | 0(0)       |
| 개정 | TOPIK II | 듣기       | 50(100)    | 1(2.0)     | 15(30.0)          | 25(50.0)          | 6(12.0)           | 3(6.0)     |
| 개정 | TOPIK II | 읽기       | 50(100)    | 2(4.0)     | 21(42.0)          | 20(40.0)          | 6(12.0)           | 1(2.0)     |

<표 III-7> 34회·35회 한국어능력시험 문항 변별도 분포

| 구분 | 시험<br>등급 | 시험<br>과목 | 문항수<br>(%) | 0.35<br>미만 | 0.35이상~<br>0.65미만 | 0.65이상~<br>1.35미만 | 1.35이상~<br>1.70미만 | 1.70<br>이상 |
|----|----------|----------|------------|------------|-------------------|-------------------|-------------------|------------|
| 기존 | 초급       | 듣기       | 30(100)    | 0(0.0)     | 1(3.3)            | 12(40.0)          | 6(20.0)           | 11(36.7)   |
| 기존 | 초급       | 읽기       | 30(100)    | 0(0.0)     | 2(6.7)            | 7(23.3)           | 10(33.3)          | 11(36.7)   |
| 기존 | 중급       | 듣기       | 30(100)    | 0(0.0)     | 1(3.3)            | 17(56.7)          | 8(26.7)           | 4(13.3)    |
| 기존 | 중급       | 읽기       | 30(100)    | 0(0.0)     | 1(3.3)            | 18(60.0)          | 8(26.7)           | 3(10.0)    |
| 기존 | 고급       | 듣기       | 30(100)    | 0(0.0)     | 2(6.7)            | 19(63.3)          | 5(16.7)           | 4(13.3)    |
| 기존 | 고급       | 읽기       | 30(100)    | 0(0.0)     | 3(10.0)           | 23(76.7)          | 4(13.3)           | 0( 0.0)    |
| 개정 | TOPIK I  | 듣기       | 30(100)    | 0(0.0)     | 1(3.3)            | 8(26.7)           | 9(30.0)           | 12(40.0)   |
| 개정 | TOPIK I  | 읽기       | 40(100)    | 0(0.0)     | 1(2.5)            | 20(50.0)          | 8(20.0)           | 11(27.5)   |
| 개정 | TOPIK II | 듣기       | 50(100)    | 0(0.0)     | 4(8.0)            | 29(58.0)          | 10(20.0)          | 4(8.0)     |
| 개정 | TOPIK II | 읽기       | 50(100)    | 0(0.0)     | 4(8.0)            | 21(42.0)          | 10(20.0)          | 15(30.0)   |

이에 본 연구에서는 문항 난이도 극단값을 연구 조건에 포함한 선행연구(김성훈, 2014; 박서홍, 이규민, 강상진, 2010; 박인용, 구슬기, 김건섭, 2013; 이현숙, 2009; 이현숙, 김성훈, 2010)와 34회·35회 한국어능력시험의 영역별 문항 난이도 결과를 바탕으로 문항 난이도 극단값이 -2, +2인 경우를 실험 조건으로 설정하였다.

#### 나. 난이도 극단값 문항의 비율

난이도 극단값을 보이는 문항의 비율을 연구조건에 포함한 선행연구(박인용, 구슬기, 김건섭, 2013)와 34회·35회 한국어능력시험의 영역별 문항 난이도 분포 결과를 바탕으로 난이도 극단값을 보이는 문항의 포함 비율이 0%, 10%, 20%인 경우를 실험 조건으로 설정하였다.

## 다. 전체 피험자 수

34회 초급 시험의 응시자는 13,293명, 34회 중급 시험의 응시자는 30,883명, 34회 고급 시험의 응시자는 16,185명으로 34회 총 응시자수는 60,361명이었다. 35회 TOPIK I 응시자는 2,367명, TOPIK II 응시자는 17,974명으로 총 20,341명이었다. 따라서 본 연구는 실제 한국어능력 시험처럼 충분한 피험자 수를 확보할 수 있는 상황을 반영하고, 안정된 결과를 산출하기 위하여 검사 X, Y의 피험자 수가 10,000명, 30,000명인 경우에 동등화 방법 간의 차이를 확인하는 연구를 진행하였다.

## 라. 가교피험자 비율

김성훈(2009)은 가교피험자 설계에서는 피험자 집단의 동등성 정도, 전체 피험자 수, 가교 피험자 비율을 고려해야 한다고 하였다. 34회 한국어능력시험 초급시험과 35회 한국어능력시험 TOPIK I 을 모두 치룬 가교피험자, 개편 전 시험인 34회 한국어능력시험 중급시험과 개편 후 시험인 35회 한국어능력시험 TOPIK II를 모두 치룬 가교피험자, 개편 전 시험인 34회 한국어능력시험 고급시험과 개편 후 시험인 35회 한국어능력시험 TOPIK II를 모두 치룬 가교 피험자 비율을 살펴보면, 34회·35회 시험의 가교피험자는 초급 169명, 중급 3,344명, 고급 1,232명이었다. 이를 바탕으로 동등화를 수행할 두 집단의 가교피험자 비율은 1%, 5%, 10%, 15%, 20%로 설정하였다.

<표 III-8> 34회·35회 한국어능력시험 공통 피험자 수

| 구분     | 34회     |         |         | 35회     |          | 공통 피험자 수 |          |
|--------|---------|---------|---------|---------|----------|----------|----------|
|        | 초급      | 중급      | 고급      | TOPIK I | TOPIK II | TOPIK I  | TOPIK II |
| 1급, 2급 | 13,293명 | -       | -       | 2,367명  | -        | 169명     | -        |
| 3급, 4급 | -       | 30,883명 | -       | -       | 17,974명  | -        | 3,344명   |
| 5급, 6급 | -       | -       | 16,185명 | -       |          | -        | 1,232명   |

본 연구에서는 검사 문항의 난이도 극단값 조건 2개, 난이도 극단값 비율 조건 3개, 전체 피험자 수(표집의 크기) 조건 2개, 가교피험자 비율 조건 5개로 총 12개 조건의 모의실험을 수행하였고, 총 5개의 동등화 방법을 사용하여 동등화 점수를 산출하였다. 각 조건별로 100회의 반복실험이 실시되었다. 따라서 전체적인 모의실험 횟수는 문항의 난이도 극단값(2) × 문항의 난이도 극단값 비율(3) × 전체 피험자 수(2) × 가교피험자 비율(5) × 동등화 방법(5) × 반복(100)으로 총 30,000회가 되었다.

### 3. 동등화 결과의 비교

동등화 표준편차는 동등화 결과의 양호성에 직접적 영향을 끼치기 때문에 상황에 적합한 설계 및 방법을 선택하는 하나의 중요한 기준으로 작용한다(Kolen & Brennan, 2004; 최성열, 이규민, 박아청, 2007). 동등화 결과의 상대적 적절성을 확인하기 위해서 동등화 이후 동등화 오차를 확인하는 과정은 필수적인 요소라고 할 수 있다. 각각의 실험 조건에 따라 수행된 동등화 방법들의 결과는 이전의 연구(Harris & Kolen, 1990; Harris & Crouse, 1993; 김현철, 2000; 이용민, 2010)에서 동등화 결과의 평가를 위한 일반적 지표들로 사용된 지표인 평균제곱오차의 제곱근(Root of Mean Squared Error; RMSE)를 사용하여 비교하였다.

평균제곱오차의 제곱근은 동등화된 점수와 동등화된 점수의 평균의 차이 제곱을 각 점수의 관측 도수로 가중평균한 제곱근의 값으로 다음 식에 의해 산출된다.

$$RMSE = \sqrt{\frac{\sum_i N_i (Eq_i - R_i)^2}{\sum_i N_i}}$$

여기서  $N_i$ 는 각 점수의 도수

$Eq_i$ 는 동등화된 점수

$R_i$ 는  $Eq_i$ 의 평균

이는 모형 또는 추정식에 의해 추정된 값과 실제 값 간의 차이를 비교하기 위해 자주 사용되는 통계량으로 값이 작다면 동등화의 결과는 그만큼 정확하고 양호도가 좋은 것을 의미한다.

### 4. 분석 도구

#### 가. 모의검사 자료의 생성

본 연구의 모의실험 자료는 실험 조건에 따라 다음과 같이 생성하였다.

1단계. R을 이용하여 문항모수  $a \sim \text{uniform}(0.5, 2.8)$ ,  $c \sim \text{uniform}(0.05, 0.3)$ 와  $b \sim N(-0.5, 1)$ 인 난수 X, Y를 생성한다.

2단계. 생성된 X, Y의 문항모수를  $\text{uniform}(0, 1)$ 에서 무선적으로 생성한 값보다 크면 1, 아니면 0을 부여한다.

3단계. 50문항에 점수 배점 2점을 곱하여 최종 점수를 산출한다.

## 나. 동등화 방법의 수행

생성된 모의실험 자료는 R을 이용하여 문항반응이론 진점수 동등화 방법(3-모수 로지스틱 동등화 방법), 선형 동등화 방법, 사전완곡화(로그선형) 동백분위 동등화 방법, 비완곡화 동백분위 동등화 방법, Circle-Arc 동등화 방법을 진행하였다. 고전검사이론 동등화 방법에 따른 동등화를 수행하기 위해 equate 패키지를 사용하였고, equate 함수에서 type으로 선형, 비완곡화 동백분위, 사전완곡화(로그 선형) 동백분위, Circle-Arc이므로 linear, equipercentile, Circle-Arc를 지정하여 동등화를 수행하였다. 두 집단을 각각 독립적인 과정으로 하여 문항모수 및 능력모수를 추정하였으며, 문항반응이론 척도변환 방법 중에서 평균-표준편차 방법을 사용하여 문항반응이론 진점수 동등화를 진행하였다.

문항반응이론 동등화 방법에 따른 동등화를 수행하기 위해 equateIRT와 ltm 패키지를 사용하였다. 문항반응이론 동등화를 수행하기 위해 equateIRT 패키지에 포함된 direc, modIRT 함수와 ltm 패키지에 내장된 tpm 함수를 사용하였다. tpm 함수에서 모수를 추정하였고, 이를 이용하여 direc, modIRT 함수에서 동등화를 수행하였다.

## 다. 동등화 결과의 비교를 위한 통계량 산출

이상의 과정을 통해 산출된 각각의 동등화 결과들은 R에 의하여 평균제곱오차의 제곱근(RMSE)이 산출되었다. 평균제곱오차의 제곱근(RMSE)을 종속변수로 하여 분산분석을 실시하였고, 이는 SAS v9.2에 의하여 수행되었다.

# IV. 분석결과

## 1. 기술 통계치

<표 IV-1>에서 검사 자료의 기술 통계치를 보면 실험 조건 별 검사 X와 검사 Y의 평균과 표준편차에서 차이를 보였다. 그 중에서 뚜렷한 차이를 보이는 것은 전체 피험자 수가 10,000명일 때 난이도 극단값이 -2인 문항이 10%, 20% 섞여 있는 경우, 검사 Y의 평균 점수가 더 높았으나, 난이도 극단값이 +2인 문항이 10%, 20% 섞여 있는 경우, 검사 X의 평균 점수가 더 높음을 알 수 있다. 그리고 전체 피험자 수가 30,000명일 때 난이도 극단값이 -2인 문항이 20% 섞여 있는 경우에는 검사 Y의 평균 점수가 더 높았으나, 난이도 극단값이 +2인 문항이 10%, 20% 섞여 있는 경우에는 검사 X의 평균이 더 높음을 알 수 있다.

<표 IV-1> 검사 자료의 기술 통계치

| 전체피험자 수  |    |       | 10,000 |        |        |        |        |        |        |        |        |        |
|----------|----|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 가교피험자 비율 |    |       | 1      |        | 5      |        | 10     |        | 15     |        | 20     |        |
| 난이도      | 비율 | 검사 유형 | 평균     | 표준 편차  | 평균     | 표준 편차  | 평균     | 표준 편차  | 평균     | 표준 편차  | 평균     | 표준 편차  |
| -2       | 0  | X     | 64.755 | 34.550 | 68.271 | 30.708 | 64.724 | 36.413 | 64.113 | 33.784 | 66.317 | 34.124 |
|          |    | Y     | 67.984 | 30.679 | 68.291 | 34.010 | 66.000 | 32.436 | 69.937 | 31.419 | 68.318 | 33.237 |
|          | 10 | X     | 67.177 | 34.332 | 71.016 | 33.259 | 67.574 | 36.185 | 67.792 | 32.159 | 68.413 | 33.584 |
|          |    | Y     | 75.400 | 29.713 | 72.593 | 32.207 | 68.373 | 33.035 | 68.839 | 33.711 | 67.105 | 31.564 |
|          | 20 | X     | 65.834 | 31.453 | 65.043 | 36.012 | 66.407 | 34.975 | 70.942 | 33.866 | 70.364 | 30.945 |
|          |    | Y     | 71.187 | 30.047 | 73.851 | 28.486 | 67.168 | 31.817 | 70.514 | 30.145 | 74.670 | 28.316 |
| +2       | 0  | X     | 64.755 | 34.550 | 68.271 | 30.708 | 64.724 | 36.413 | 64.113 | 33.784 | 66.317 | 34.124 |
|          |    | Y     | 67.984 | 30.679 | 68.291 | 34.010 | 66.000 | 32.436 | 69.937 | 31.419 | 68.318 | 33.237 |
|          | 10 | X     | 68.089 | 32.897 | 66.715 | 32.257 | 68.506 | 33.091 | 64.254 | 37.828 | 64.952 | 34.107 |
|          |    | Y     | 67.404 | 31.288 | 64.684 | 33.253 | 66.671 | 29.471 | 63.925 | 31.681 | 60.426 | 31.270 |
|          | 20 | X     | 71.837 | 31.673 | 71.533 | 32.062 | 71.610 | 31.963 | 71.864 | 31.750 | 72.060 | 31.580 |
|          |    | Y     | 56.666 | 31.806 | 55.709 | 32.119 | 55.831 | 31.989 | 55.927 | 31.642 | 56.024 | 31.880 |
| 전체피험자 수  |    |       | 30,000 |        |        |        |        |        |        |        |        |        |
| 가교피험자 비율 |    |       | 1      |        | 5      |        | 10     |        | 15     |        | 20     |        |
| 난이도      | 비율 | 검사 유형 | 평균     | 표준 편차  | 평균     | 표준 편차  | 평균     | 표준 편차  | 평균     | 표준 편차  | 평균     | 표준 편차  |
| -2       | 0  | X     | 65.482 | 35.607 | 63.408 | 35.856 | 69.002 | 35.013 | 65.230 | 34.308 | 67.045 | 33.687 |
|          |    | Y     | 68.110 | 31.862 | 67.464 | 34.021 | 68.889 | 33.542 | 68.297 | 32.981 | 67.345 | 33.604 |
|          | 10 | X     | 71.430 | 33.024 | 66.505 | 33.021 | 68.739 | 35.373 | 68.490 | 33.799 | 65.110 | 34.855 |
|          |    | Y     | 72.249 | 33.225 | 64.786 | 31.992 | 67.326 | 34.045 | 65.467 | 33.561 | 70.553 | 29.796 |
|          | 20 | X     | 64.167 | 33.772 | 65.519 | 36.614 | 68.318 | 35.562 | 67.405 | 32.354 | 69.386 | 34.988 |
|          |    | Y     | 72.705 | 28.599 | 72.620 | 30.158 | 70.312 | 31.757 | 73.039 | 31.548 | 70.052 | 31.758 |
| +2       | 0  | X     | 65.482 | 35.607 | 63.408 | 35.856 | 69.002 | 35.013 | 65.230 | 34.308 | 67.045 | 33.687 |
|          |    | Y     | 68.110 | 31.862 | 67.464 | 34.021 | 68.889 | 33.542 | 68.297 | 32.981 | 67.345 | 33.604 |
|          | 10 | X     | 67.759 | 32.656 | 70.255 | 32.286 | 66.568 | 34.682 | 68.759 | 33.487 | 72.304 | 32.602 |
|          |    | Y     | 62.742 | 31.282 | 63.079 | 35.113 | 66.460 | 31.462 | 65.914 | 31.321 | 61.939 | 33.266 |
|          | 20 | X     | 71.953 | 31.745 | 71.699 | 31.934 | 71.956 | 31.752 | 71.605 | 31.836 | 71.753 | 31.897 |
|          |    | Y     | 55.792 | 31.950 | 55.699 | 31.892 | 55.797 | 31.939 | 55.678 | 32.110 | 56.036 | 31.951 |

## 2. 동등화 오차 통계량

<표 IV-2>의 동등화 오차(RMSE)에 대한 분산분석 결과에 의하면, 유의수준이 .05일 경우 주효과가 유의한 조건은 동등화 방법(E), 문항의 난이도 극단값(D), 문항의 난이도 극단값 비율(R), 공통 피험자 비율(P)이었다. 상호작용 효과가 유의한 조건은 동등화 방법(E)과 문항의 난이도 극단값(D), 동등화 방법(E)과 문항의 난이도 극단값 비율(R), 문항의 난이도 극단값(D)과 문항의 난이도 극단값 비율(R), 문항의 난이도 극단값(D)과 전체 피험자 수(N), 문항의 난이도 극단값(D)과 공통 피험자 비율(P), 문항의 난이도 극단값 비율(R)과 전체 피험자 수(N), 문항의 난이도 극단값 비율(R)과 공통 피험자 비율(P), 전체 피험자 수(N)과 공통 피험

자 비율(P)임을 알 수 있다.

<표 IV-2> RMSE에 대한 분산분석 결과

| Source            | DF  | SS      | MS      | F       | P      |
|-------------------|-----|---------|---------|---------|--------|
| 동등화 방법(E)         | 4   | 809.36  | 202.34  | 196.17  | <.0001 |
| 문항의 난이도 극단값(D)    | 1   | 1718.47 | 1718.47 | 1666.11 | <.0001 |
| 문항의 난이도 극단값 비율(R) | 2   | 87.73   | 43.87   | 42.53   | <.0001 |
| 전체 피험자 수(N)       | 1   | 3.63    | 3.63    | 3.52    | 0.0623 |
| 가교피험자 비율(P)       | 4   | 17.62   | 4.41    | 4.27    | 0.0025 |
| E×D               | 4   | 191.82  | 47.95   | 46.49   | <.0001 |
| E×R               | 8   | 21.32   | 2.67    | 2.58    | 0.0107 |
| E×N               | 4   | 2.60    | 0.65    | 0.63    | 0.6407 |
| E×P               | 16  | 4.25    | 0.27    | 0.26    | 0.9984 |
| D×R               | 2   | 156.92  | 156.92  | 152.14  | <.0001 |
| D×N               | 1   | 12.61   | 12.61   | 12.23   | 0.0006 |
| D×P               | 4   | 18.07   | 4.52    | 4.38    | 0.0021 |
| R×N               | 2   | 7.42    | 3.71    | 3.60    | 0.0293 |
| R×P               | 8   | 21.93   | 2.74    | 2.66    | 0.0088 |
| N×P               | 4   | 16.07   | 4.02    | 3.90    | 0.0046 |
| 오차                | 185 | 190.81  | 1.03    |         |        |

다음 <표 IV-3>와 <표 IV-4>에는 분산분석에 의하여 유의한 것으로 판정된 동등화 방법들과 조건 간의 상호작용 효과 해석을 위하여 RMSE 크기가 제시되었다. <표 IV-3>에는 동등화 방법(E)과 문항의 난이도 극단값(D)의 상호작용 결과가 제시되었다. 이 표에 의하면 문항의 난이도 극단값 조건에서 난이도 극단값이 +2일 때가 난이도 극단값이 -2일 때보다 오차가 매우 컸다. 이는 두 개의 검사의 난이도 차이가 주원인으로 보인다. 난이도 극단값의 모든 조건에서 Circle-Arc 동등화 방법의 오차가 가장 작았고, 사전완곡화(로그 선형) 동백분위 동등화 방법의 오차가 가장 컸음을 알 수 있다.

<표 IV-3> RMSE에 대한 동등화 방법(E)과 난이도 극단값(D)의 상호작용

| 동등화 방법<br>난이도 극단값 | IRT 진점수 | 선형     | 로그선형<br>동백분위 | 비완곡화<br>동백분위 | Circle-<br>Arc |
|-------------------|---------|--------|--------------|--------------|----------------|
| -2                | 4.6970  | 2.4849 | 5.5045       | 4.9170       | 1.6422         |
| +2                | 7.4383  | 8.4053 | 12.5783      | 11.6736      | 6.0375         |

<표 IV-4>에는 동등화 방법(E)과 문항의 난이도 극단값 비율(R)의 상호작용 결과가 제시되었다. 이 표에 의하면 문항의 난이도 극단값 비율이 커질수록 IRT 진점수 방법, 선형 방법, Circle-Arc 방법은 오차가 커졌으나, 비완곡화 동백분위 방법과 사전완곡화(로그 선형) 동백분위 방법은 난이도 극단값 문항의 포함 비율이 20%일 때보다 10%일 때의 오차가 더 컸다. 난이도 극단값 비율의 모든 조건에서 Circle-Arc의 오차가 가장 작았고, 사전완곡화(로그 선형) 동백분위 방법의 오차가 가장 컸음을 알 수 있다.

<표 IV-4> RMSE에 대한 동등화 방법(E)과 난이도 극단값 비율(R)의 상호작용

| 난이도 극단값 비율 \ 동등화 방법 | IRT<br>진점수 | 선형     | 로그선형<br>동백분위 | 비완곡화<br>동백분위 | Circle-<br>Arc |
|---------------------|------------|--------|--------------|--------------|----------------|
| 0%                  | 4.3087     | 3.7574 | 7.2652       | 6.1852       | 2.5797         |
| 10%                 | 5.9080     | 4.9765 | 9.1653       | 8.1730       | 3.5221         |
| 20%                 | 6.4215     | 5.2775 | 8.0371       | 7.7834       | 3.6889         |

다음 <표 IV-5>부터 <표 IV-10>에는 분산분석의 결과가 유의한 것으로 나타나는 실험 조건 간의 상호작용 효과 해석을 위하여 RMSE 크기가 제시되었다. <표 IV-5>에는 문항의 난이도 극단값(D)과 문항의 난이도 극단값 비율(R)의 상호작용 결과가 제시되었다. 이 표에 의하면 문항의 난이도 극단값 -2가 포함된 조건에서는 난이도 극단값 포함 비율이 커질수록 오차가 감소하였으나, 문항의 난이도 극단값 +2가 포함된 조건에서는 난이도 극단값 포함 비율이 커질수록 오차가 커졌다. 이에 대한 원인은 응시자의 능력이 -2 또는 +2에 가까운 분포로 인한 오차 발생이거나, 가교 피험자의 능력이 평균보다 낮았던 것이 원인으로 작용하였던 것으로 보인다.

<표 IV-5> RMSE에 대한 난이도 극단값(D)과 난이도 극단값 비율(R)의 상호작용

| 난이도 극단값 \ 난이도 극단값 비율 | 0%     | 10%    | 20%     |
|----------------------|--------|--------|---------|
| -2                   | 4.8192 | 4.3035 | 2.4246  |
| +2                   | 4.8192 | 8.3945 | 10.0587 |

<표 IV-6>에는 문항의 난이도 극단값 조건(D)과 전체 피험자 수 조건(N)의 상호작용 결과가 제시되었는데, 전체 피험자 수 모든 조건에서 난이도 극단값이 +2일 때가 난이도 극단값이 -2일 때보다 오차가 매우 컸다. 난이도 극단값 조건에서 전체 피험자 수가 30,000명일 때가 10,000명일 때보다 오차가 감소하였다.

&lt;표 IV-6&gt; RMSE에 대한 난이도 극단값(D)과 전체 피험자 수(N)의 상호작용

| 난이도 극단값 \ 전체 피험자 수 | 10,000명 | 30,000명 |
|--------------------|---------|---------|
| -2                 | 4.1261  | 3.5721  |
| +2                 | 9.3598  | 9.0934  |

<표 IV-7>에는 문항의 난이도 극단값 조건(D)과 가교피험자 비율 조건(P)의 상호작용 결과가 제시되었는데, 가교피험자 비율의 모든 조건에서 난이도 극단값이 +2일 때가 난이도 극단값이 -2일 때보다 오차가 매우 컸다. 난이도 극단값이 -2일 경우에는 가교피험자 비율이 1%일 때 오차가 가장 컸으며, 난이도 극단값이 +2일 경우에는 가교피험자 비율 조건에서 큰 차이를 보이지 않았다.

&lt;표 IV-7&gt; RMSE에 대한 난이도 극단값(D)과 가교피험자 비율(P)의 상호작용

| 난이도 극단값 \ 가교피험자 비율 | 1%     | 5%     | 10%    | 15%    | 20%    |
|--------------------|--------|--------|--------|--------|--------|
| -2                 | 4.6982 | 3.6315 | 3.6368 | 3.6310 | 3.6481 |
| +2                 | 9.2776 | 9.2021 | 9.2787 | 9.1549 | 9.2197 |

<표 IV-8>에는 문항의 난이도 극단값 비율(R)과 전체 피험자 수 조건(N)의 상호작용 결과가 제시되었는데, 전체 피험자 수 모든 조건에서 난이도 극단값 비율이 커질수록 오차가 컸으며, 난이도 극단값 조건에서 전체 피험자 수가 30,000명일 때 10,000명일 때보다 오차가 감소하였다.

&lt;표 IV-8&gt; RMSE에 대한 난이도 극단값 비율(R)과 전체 피험자 수(N)의 상호작용

| 난이도 극단값 비율 \ 전체 피험자 수 | 10,000명 | 30,000명 |
|-----------------------|---------|---------|
| 0%                    | 4.8744  | 4.7641  |
| 10%                   | 6.7661  | 5.9319  |
| 20%                   | 6.3457  | 6.1377  |

<표 IV-9>를 보면 난이도 극단값 비율이 0%일 때 가교피험자 비율의 모든 조건에서 오차가 가장 작았다. 그리고 난이도 극단값 비율이 20%일 때 가교피험자 비율의 대부분 조건에서 오차가 가장 컸으나, 난이도 극단값 비율이 10%이고, 가교피험자 비율이 1%일 때의 경우에만 난이도 극단값 비율이 20%일 때에 비하여 오차가 컸다.



<표 IV-9> RMSE에 대한 난이도 극단값 비율(R)과 가교피험자 비율(P)의 상호작용

| 가교피험자 비율<br>난이도극단값 비율 | 1%     | 5%     | 10%    | 15%    | 20%    |
|-----------------------|--------|--------|--------|--------|--------|
| 0%                    | 4.8628 | 4.8473 | 4.7802 | 4.8249 | 4.7809 |
| 10%                   | 7.6018 | 6.0241 | 6.0544 | 6.0211 | 6.0434 |
| 20%                   | 6.2917 | 6.2016 | 6.2893 | 6.1678 | 6.2580 |

<표 IV-10>을 보면 전체 피험자 수가 10,000명일 때와 30,000명 일 때 모두 공통 피험자 비율이 1%일 때 오차가 가장 컸으며, 5%, 10%, 15%, 20%일 때의 오차는 비슷하였다. 그리고 전체 피험자 수가 30,000명일 때는 10,000명일 때보다 가교피험자 비율 조건에서 모두 오차가 작았다.

<표 IV-10> RMSE에 대한 전체 피험자 수(N)와 가교피험자 비율(P)의 상호작용

| 가교피험자 비율<br>전체 피험자 수 | 1%     | 5%     | 10%    | 15%    | 20%    |
|----------------------|--------|--------|--------|--------|--------|
| 10,000명              | 7.2311 | 5.9692 | 6.0237 | 5.9210 | 5.9530 |
| 30,000명              | 5.8288 | 5.7503 | 5.7634 | 5.7601 | 5.8004 |

## V. 결론 및 논의

본 연구를 통해 얻은 결과를 실험조건별로 제시하면 다음과 같다. 난이도 극단값 조건과 난이도 극단값 문항의 포함 비율 조건에 대해 살펴보면 모든 실험 조건에서 난이도 극단값이 +2일 때가 난이도 극단값이 -2일 때보다 오차가 매우 컸고, 난이도 극단값 -2가 포함된 조건에서는 난이도 극단값 포함 문항의 비율이 커질수록 오차가 감소하였으나, 문항의 난이도 극단값 +2가 포함된 조건에서는 난이도 극단값 포함 비율이 커질수록 오차가 커졌다. 그리고 난이도 극단값 문항의 포함 비율이 0%일 때 가교피험자 비율의 모든 조건에서 오차가 작았고, 난이도 극단값이 -2일 경우에는 가교피험자 비율이 1%일 때 오차가 가장 컸으며, 난이도 극단값이 +2일 경우에는 모든 가교피험자 비율 조건에서 큰 차이를 보이지 않았다. 전체 피험자 수 조건은 10,000명과 30,000명으로 설정하였는데, 모든 실험 조건에서 전체 피험자 수가 30,000명일 때는 10,000명일 때보다 오차가 감소하였다. 전체 피험자 수를 1,000명, 10,000명, 100,000명으로 설정한 Godfrey(2007)는 표본의 크기가 증가하면 모든 동등화 방법들이 유사한 결과를 나타냈다고 하였다. 그리고 가교피험자 비율 조건에 대해 살펴보면, 전체 피험자 수의 1%일 때 오차가 가장 컸으며, 5%, 10%, 15%, 20%일 때의 오차는 비슷하였다. 또한 동

등화 방법별로 살펴보면, 전체 실험 조건에서 Circle-Arc 동등화 방법의 오차가 가장 작았고, 로그선형 사전완곡화 동백분위 동등화 방법의 오차가 가장 컸다. Livingston & Kim(2009)은 Circle-Arc 동등화 방법에서는 최저점의 위치가 달라짐에 따라 동등화 오차가 감소한다고 하였다. 즉, Circle-Arc 동등화 방법에서 최저점의 위치가 0부터 1씩 증가함에 따라 동등화 오차는 계속 감소하였다. 본 연구의 경우, 50문항 4지 선다형이므로 추측요인으로 정답을 선택할 수 있는 최저점의 위치가 증가한 것이 동등화 오차 감소의 요인으로 작용한 것으로 보인다. 본 연구에서 가장 동등화 오차가 큰 동등화 방법은 동백분위 동등화 방법과 로그선형 사전 완곡화 동등화 방법이었는데, 우선 동백분위 동등화 방법은 점수의 간격이 1점일 때 동등화 점수의 최저점이  $-0.5$ 점보다 작지 않으며, 최고점이 만점  $+0.5$ 점보다 크지 않도록 하는 원칙에 의하여 영점과 만점 부근에서 실험 조건을 일관성 있게 반영한 동등화 점수를 산출하지 못하였다. Cook & Eignor(1983)는 GRE 등의 시험을 동등화하는 연구에서 동백분위 동등화 방법은 극단점수에서의 자료가 적을 경우 적합하지 않다는 결과를 보여주었다. 하지만 김현철(2003)은 난이도가 서로 다른 두 개의 검사를 동일한 집단에게 실시할 때 일반적으로 두 검사에서의 점수 변화는 중간 정도의 학업 능력을 가진 집단과 상위 또는 하위의 학업능력을 가진 집단에서 동일한 기울기를 가진 한 개의 선형변환에 의해서 설명되지 않는다고 하였다. 한 시험에서 영점에 가까운 점수를 획득한 피험자는 난이도가 더 쉬운 검사에서도 난이도의 차이만큼 일정하게 증가한 점수보다는 영점에 더 가까운 점수를 획득할 가능성이 많으며, 만점에 가까운 점수를 획득한 피험자는 난이도가 더 어려운 검사에서도 난이도의 차이만큼 일정하게 감소한 점수보다는 만점에 더 가까운 점수를 획득할 가능성이 많다. 따라서 동백분위 동등화 방식을 사용하는 동등화 방법들의 동등화 결과가 선형 동등화 방식을 사용하는 동등화 방법들의 동등화 결과와는 달리 영점과 만점 부근에서 실험 조건의 차이에 비례한 점수변환이 일어나지 않는 것은 오히려 적절한 동등화의 결과일 수 있다고 주장하였다. Shin(2011)은 무선집단설계에서 로그선형 사전완곡화 동백분위 방법이 다양한 자료의 조건에서 일반적으로 각각의 원점수에 더 나은 통계적 적합성을 나타냈다고 평가하였지만 본 연구에서는 로그선형 사전완곡화 동백분위 동등화 방법의 동등화 오차가 가장 컸다.

본 연구처럼 검사에 난이도가 극단적인 문항이 포함될 경우에는 동등화를 수행 시 매우 신중을 기하여야 하는데, 이에 대해 최태진(1991)은 여러 검사형 간에 문항난이도 평균 및 표준편차의 차이가 클 경우에는 검사의 신뢰도 하락 및 일차원성의 이탈을 수반한다고 밝혔다. 그리고 이종승(1993)도 어떠한 동등화 방법을 적용하건 동등화 점수에는 거의 언제나 전환에 따른 오차가 개재될 가능성이 있으므로 검사의 특성과 검사 집단의 능력수준이 서로 현격한 차이를 보일 때 이러한 차이점들을 모두 수용하면서 동등화할 수 있는 최선의 방법은 없으므로 극히 이질적인 검사자료는 동등화하지 않는 것이 바람직하다고 밝혔다. 박인용, 구슬기, 김진섭(2013)은 공통문항 비동등집단 설계에서 고정문항모수 추정방법을 통해 공통적으로 변환검

사를 추정할 때 극단 정도가 -4인 문항의 비율이 10%이상일 경우 오차가 크게 나타나며, 변환검사의 특성곡선이 약 1점 이상의 과대 혹은 과소추정이 발생하는 것을 확인하였다. 그리고 이현숙(2009)은 동등화에 사용된 두 검사의 비공통 문항 난이도 차이가 클수록, 그리고 두 검사를 치른 집단 간의 능력 차이가 클수록 이러한 패턴은 더욱 뚜렷하게 관측되었다고 밝혔다. 즉 검사에 난이도가 극단적인 문항이 포함될 경우와 검사형 간에 문항, 집단의 특성이 다를 경우에는 동등화 수행 시 매우 신중을 기하여야 할 것이다.

본 연구에서 고려된 조건들 외에 다양한 조건의 검사 상황을 고려한 추가적인 연구들을 통해 여러 동등화 방법들과 여러 실험 조건들을 비교하여 그 차이점들을 밝혀낼 수 있다면, 가교피험자 설계에서 실제로 발생할 수 있는 다양한 조건의 검사들에 알맞은 최적의 동등화 방법을 선택할 수 있는 데 도움이 될 수 있을 것으로 본다.

## 참 고 문 헌

- 김성훈(2009). IRT 공통 능력척도 개발을 위한 공통-피험자 척도연계 방법과 공통-문항 척도연계 방법의 상대적 기능 연구, **교육평가연구**, 22(4), 1057-1077.
- 김성훈(2014). 동등집단 설계를 사용하는 수직척도화에서 공통 능력척도의 개발을 위한 IRT 추정 방법의 비교, **교육평가연구**, 27(4), 925-944.
- 김현철(2000). 순환동등화에 의한 2000학년도 대입수능시험 선택과목 점수 동등화결과의 비교, **교육평가연구**, 13(2), 153-168.
- 김현철(2003). 모의실험에 의한 대학수학능력시험 선택과목 점수 동등화 방법들의 적합성 비교 (I): 원점수와 동등화점수의 차이분석, **교육학연구**, 41(1), 363-384.
- 남현우(2001). **검사동등화 방법**. 교육과학사.
- 민경석(2014). **한국어능력시험 개편 체제 점수 체계 적용 방안 연구**. 국립국제교육원.
- 박서홍·이규민·강상진(2010). A Comparison of Kernel and Traditional Equating Methods under the Non-Equivalent Groups with Anchor Test Design, **교육평가연구**, 23(4), 975-1002.
- 박인용·구슬기·김건섭(2013). 고정문항모수 추정방법에서 고유문항의 난이도 극단 특성이 척도변환에 미치는 영향, **교육평가연구**, 26(1), 107-129.
- 안수현(2016). 모의실험에 의한 가교피험자설계의 조건에 따른 동등화 방법 비교. 박사학위논문, 성균관대학교.
- 이상하·박도영(2012). 문항 예측정답률과 문항반응이론을 이용한 원점수 분포 추정, **교육평가연구**, 25(3), 459-477.
- 이용민(2010). 모의실험에 의한 커널 동등화 방법과 다른 동등화 방법들의 비교. 박사학위논문, 성균관대학교.
- 이종승(1993). 검사점수의 동등화, **교육평가연구**, 6(1), 169-196.
- 이현숙(2009). Evaluating the performance of non-equivalent groups anchor test equating under various conditions of anchor test construction. **교육평가연구**, 22(3), 847-869.
- 이현숙·김성훈(2010). 외적 가교검사의 통계적 구성 조건 완화가 검사동등화 결과에 미치는 영향, **교육평가연구**, 23(2), 417-439.
- 최성열·이규민·박아청(2007). 검사점수 연계화 방법의 상대적 적절성 비교, **교육심리연구**, 21(1), 1-25.
- 최태진(1991). 검사점수 동등화를 위한 동등백분위법, 선형법, 문항반응이론 기법의 비교. 석사학위논문, 부산대학교.

- Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. *Applications of Item Response Theory*, Vancouver: Educational Research Institute of British Columbia, 175-195.
- Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.
- Godfrey, K. E. (2007). A comparison of kernel equating and IRT true score equating methods (Doctoral dissertation, University of North Carolina).
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*, 7. Springer Science & Business Media.
- Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement*, 15(4), 391-408.
- Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement*, 50(1), 61-71.
- Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6(3), 195-240.
- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, 25(2), 133-183.
- Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, 22, 197-206.
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and Practices*. New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.
- Livingston, S. A., & Kim, S. (2009). The Circle-Arc Method for Equating in Small Samples. *Journal of Educational Measurement*, 46(3), 330-343.
- Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory*. ETS Research Bulletin Series, 1975(1), i-35.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
- R Core Team. (2013). R: A language and environment for statistical computing

- [Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved October 1, 2014, from <http://www.R-project.org>.
- Segall, D. O. (1997). *Equating the CAT-ASVAB*. In W. A. Sands, Computerized adaptive testing: From inquiry to operation (pp. 181-198). Washington, DC: American Psychological Association.
- Shin, S. (2011). A Comparison of Van der Linden's conditional equipercentile equating method with other equating methods under the random groups design (Doctorial dissertation, University of Iowa).
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.
- Wang, C. (2011). An investigation of bootstrap methods for estimating the standard error of equating under the common-item nonequivalent groups design (Doctorial dissertation, University of Iowa).
- 한국어능력시험 홈페이지 <http://www.topik.go.kr>(검색일: 2016. 07. 08.)

· 논문접수 : 2016.12.20. / 수정본접수 : 2017.02.13. / 게재승인 : 2017.02.14.

## ABSTRACT

### A Comparison of Equating Methods on Various Conditions in Anchor Persons Design by Simulation

Su-Hyun Ahn

Researcher, National Institute for International Education

Hyun-Chul Kim

Professor, Sungkyunkwan University

The purpose of this study is to provide guidelines for choosing test equating method on specific test conditions. For this purpose, this study investigates the differences between equating methods on various conditions in anchor persons design by simulation. In creating the material, the three parameter logistic model for the multiple-choice items were used. The parameters for the three parameter logistic model selected 50 items in  $a \sim \text{uniform}(0.5, 2.8)$ ,  $b \sim N(-0.5, 1)$ ,  $c \sim \text{uniform}(0.05, 0.3)$ . For the range of the ability of the subjects,  $N(0, 1)$  for base group was set up. To make up the condition for the equating, five factors( $2 \times 3 \times 2 \times 5 \times 5$ ), including the degree of extreme(-2, +2), the proportions of items with extreme difficulty(0%, 10%, 20%), the sample size(10,000, 30,000), the proportions of anchor persons(1%, 5%, 10%, 15%, 20%), the equating methods (IRT true score equating method, linear equating method, non-smoothing equipercentile equating method, log-linear presmoothing equipercentile equating method, Circle-Arc equating method) were taken into consideration and a total of 300 conditions were examined in this study. To compare equating methods, the generated data were equated using R programing code by the author. Additionally the results from all equating methods were compared by RMSE, MSD, MAD with 100 replications. Also, to analyze the effects of conditions included, the equated scores by all equating methods were analyzed by ANOVA using SAS 9.2.

*Key Words : Simulation, Anchor Persons Design, CTT equating method, IRT equating method, test equating, equating error*