교육과정평가연구
The Journal of Curriculum and Evaluation
2009, Vol. 12, No. 2, pp. 51～72

사례수 차이가 검사동등화의 집단불변성 원칙 검증에 미치는 영향

김 희 경(한국교육과정평가원 부연구위원)

≪ 요 약 ≫

집단불변성(population invariance)이란 검사동등화에 있어 필수 불가결한 원칙이다. 이 원칙에 따르면 동등화 결과는 어떤 집단을 사용하는지에 상관없이 항상 유일해야한다. 성별, 인종, 능력, 그밖에 여러 가지 중요한 특성에 따라 구분된 집단들이 검사동등화의 집단불변성 원칙을 충족하는지 검증하는 많은 선행 연구들이 수행되었다. 이러한 집단불변성 원칙 관련 연구에서 집단 간 동등화 차이를 수치로 나타내기 위해 REMSD(Root Expected Mean Square Difference)가 빈번히 계산되었다. 하지만 비교하려는 집단들 간 사례수가 크게 차이나는 경우 REMSD값은 작아지는 경향이 있고 따라서 집단불변성 원칙의 평가가 실제와 다르게 판단될 수 있다는 우려가 제기되었다. 따라서 REMSD와 같은 오차 통계치를 판단 기준으로 사용하여 집단불변성 원칙을 평가하고자 할 때 사례수의 변화에 따라 얼마나 민감하게 영향을 받는지 검증할 필요가 있다. 본 연구는 집단의 사례수가 REMSD 계산에 실제로 얼마나 민감한 영향을 끼치는가를 분석하였다. 또한 REMSD 계산 과정에서 각 집단에 부과되는 가중치(weight)를 결정할 때 여러 가지 방법을 사용하여 가중치를 부여하여 사례수에 보다 덜 민감한 REMSD값을 구할 수 있는지 조사하였다. 연구 결과, 집단 간 사례수 차이에 따라 오차 통계치(RMSD)의 값이 원래보다 커지거나 작아지기도 하는 것을 확인하였다. 따라서 비교하려는 집단 간 크기가 서로 비슷하지 않고 불균형한 경우 REMSD값을 사용하여 집단불변성 원칙을 평가하고자 하는 것에 대해 유의해야 할 것이며, 사례수의 영향을 받지 않는 판단기준을 개발해야 할 필요성을 강조하였다.

주제어 : 검사동등화, 집단불변성 원칙, 동백분위 동등화, REMSD, RMSD

Ⅰ. 서론

검사동등화가 성공적으로 수행되었다고 판단되는 경우는 대칭성(symmetry), 공정성(equity), 집단불변성(population invariance) 원칙 등이 지켜진 경우를 말한다(Kolen & Brennan, 2004).

즉, 집단불변성 원칙은 검사동등화의 필수 불가결한 원칙으로서 이 원칙에 따르면 동등화 결과는 동등화에 사용된 집단에 상관없이 항상 유일무이해야 한다. 검사동등화가 시행되는 목적은 동형 검사 간 난이도 차이를 통계적으로 조정해줌으로써 피험자가 어떤 유형의 검사를 치렀는지에 상관없이 점수 간의 공정한 비교를 할 수 있도록 하려는 것이다. 따라서 동등화 결과로 산출된 원점수에서 척도점수로의 전환표는 어떤 집단에서든 동일하게 적용될 수 있어야 한다. 성별 집단을 예로 든다면, 남자 집단을 이용한 동등화 결과와 여자 집단을 이용한 동등화 결과는 항상 일치해야 한다. 이러한 집단불변성 원칙은 검사동등화가 성공적으로 수행되었는지 여부를 검증하는 도구가 될 수 있다. 지금까지 집단불변성 원칙을 검증하여 동등화의 양호성을 평가하는 많은 연구들이 이루어졌다. 성별 이외에도 능력, 인종, 전공, 모국어의 종류, 관련 과목의 이수 여부에 의해 구분된 집단 간에 집단불변성 원칙이 지켜지는지 검증하는 많은 선행 연구들이 수행되었다(Angoff & Cowell, 1986; Harris & Kolen, 1986; Dorans & Holland, 2000; Tateneni & Dorans, 2003; Yi et al., 2004).

미국에서는 대학입학 시험이나 면허 또는 자격증을 취득하기 위한 여러 종류의 시험에서 동등화가 사용되고 있다. 검사 점수가 개인에게 끼치는 영향력이 큰 중요한 시험일수록 점수의 공정성을 높이고자 검사동등화를 실시하고 있다. 한국의 경우에도 국가수준 학업성취도 평가에서 연도별 변화 추이를 살펴보고자 동등화가 도입되어 시행되고 있다. 한국의 수능이나 국가수준의 임용고시 등에서 동등화가 사용되는 것은 여러 제약점에 의해 간단한 일이 아니지만 각종 면허 또는 자격증 시험에 검사동등화 이용이 증가하는 것을 예상해 볼 수 있다.

집단불변성 원칙은 일반적인 수평적 검사동등화뿐 아니라 수직적 점수연계(vertical scaling)에 있어서 그 양호성을 평가하는데 사용될 수 있고, 또한 ACT와 SAT간 점수연계와 같이 서로 다른 종류의 검사 간 점수를 비교하고자 점수를 일치화(concordance)하는 경우에 있어서도 그 양호성을 평가하기 위해 사용될 수 있는 중요한 특성이다. 수직적 점수연계는 개인이나 학교, 시도교육청 등의 시간에 따른 학력 성장을 탐색하기 위해 유용하다. 미국은 NCLB(No Child Left Behind)의 시행에 따라 학교 효과를 평가하기 위하여 연도별로 각 학교의 학력 성장이 목적하는 만큼 이루어지고 있는지 확인하고 있는데, 이에 따라 수직적 점수 연계의 중요성이 증가하였다. 또한 서로 다른 기관에서 개발한 비슷한 목적의 검사 간에 점수를 비교해야 하는 상황이 요구될 수도 있다. 미국의 ACT와 SAT 간 점수를 비교하기 위해 점수 일치화(concordance)를 하는 것을 예를 들 수 있는데, 학생, 학부모, 대학 입학사정 관리자들이 유용한 참고자료로 이용할 수 있도록 제공되고 있다. 집단불변성 원칙은 이러한 검사의 수평적 동등화, 수직적 연계화, 또는 비슷한 성격의 검사 점수 일치화에서, 그 수행의 양호도를 보장하기 위해 지켜져야 하는 바람직한 원칙이므로 이 원칙을 올바르게 검증해야 할 필요가 있다.

집단불변성 원칙을 평가하기 위해서는 관심의 대상이 되는 집단(예를 들어, 남자, 여자)을 정하여, 각 집단별로 따로 동등화/연계화/일치화를 실시한 후 집단에 따라 결과에 차이가 있는지 비교한다. 이때 차이가 어느 정도 큰지 판단하기 위해서 주로 오차 통계치를 계산하여 판단의 기준으로 삼는다. Dorans & Holland(2000)가 집단불변성 원칙을 평가하기 위한 오차 통계치로서 REMSD(Root Expected Mean Square Difference)를 사용할 것을 제안한 후 많은 집단불변성 연구에서 REMSD를 계산하여 집단 간 동등화 차이의 크기를 판단하였다(Tateneni & Dorans, 2003; Dorans, 2004; Yang, 2004; Yi et al., 2004; von Davier & Han, 2004; von Davier et al., 2004; von Davier & Wilson, 2004). 하지만 REMSD의 부주의한 사용에 대해 경고가 제기되었다. Yang et al.(2003) 및 Yang(2004)은 집단의 사례수에 따라 RMSD값이 크게 영향을 받는다는 사실을 보고하였다. 그들 연구에 따르면 비교하려는 집단 간에 사례수가 비슷하지 않고 불균형한 경우 REMSD값은 작아지는 경향이 있고, 작아진 REMSD값에 의해 실제로 그렇지 못한 경우에도 집단불변성 원칙이 지켜진다고 잘못 평가될 가능성이 있다고 하였다. 요컨대 사례수의 영향에 의해 집단불변성 원칙에 대한 왜곡된 평가가 내려질 수 있다는 것이다. 집단의 사례수가 불균형 한 경우, REMSD에 어떠한 영향을 미치는지 정확히 파악하는 것은 집단 불변성 원칙 검증 결과를 해석하는데 도움이 될 것이다.

이 연구에서는 두 집단 간 사례수 차이가 1:1, 2:1, 3:1, 4:1, 5:1로 변하는 경우 검사동등화의 집단불변성을 REMSD를 계산하여 확인해 본다. 집단의 크기가 불균형할 때 REMSD값이 작아진다는 선행 연구의 결과를 검증하기 위해 집단 간 사례수 차이가 점점 커짐에 따라 REMSD값이 얼마나 민감한 영향을 받는지 탐색해 보고자 한다. 또한 REMSD 계산 과정에서 편차점수의 제곱(squared difference)을 평균내기 위해 사용되는 가중치(weight)를 네 가지의 서로 다른 방법으로 계산해본다. 가중치를 부여하는 방식에 따라 REMSD가 사례수에 덜 영향을 받도록 개선할 수 있는지 논의할 것이다.

구체적으로 본 연구에서 조사하고자 하는 연구문제는 다음의 두 가지이다.

1) 두 집단 간 사례수가 1:1, 2:1, 3:1, 4:1, 5:1로 변함에 따라 REMSD값이 어떻게 변하는가?
2) REMSD 계산 과정에서 가중치를 부여하는 방법을 네 가지로 달리 적용해 보았을 때 어떠한 차이가 있는가?

Ⅱ. 이론적 배경

1. REMSD의 계산

Dorans & Holland(2000)가 집단 간 동등화 차이를 수치화하기 위해 REMSD를 사용할 것을

제안하였다. REMSD를 계산하기 위해서는 먼저 RMSD(Root Mean Squared Difference)를 계산한다. 우선 두 가지 동형검사 X와 Y에서 관찰된 점수를 각각 x, y로 표시하기로 하고, 검사 X가 검사 Y로 동등화된다고 가정한다. 검사 X를 치른 모든 피험자를 포함하는 전체 집단 P에 속해있는 관심의 대상이 되는 하위 집단들을 P_1, P_2, P_3, …, P_j 라고 표시하기로 할 때 RMSD의 계산 공식은 다음과 같다.

$$RMSD(x)=\frac{\sqrt{\sum_j w_j[S_{p_j}(x)-S_p(x)]^2}}{\sigma_P(Y)}. \quad (1)$$

위식에서 P(전체 집단)와 P_j(j번째 하위 집단)에서 동등화 결과로 성립된 원점수-척도점수 관계를 나타내는 함수를 각각 $S_p(x)$와 $S_{p_j}(x)$로 표시하였다. RMSD의 계산 과정에서는 각 하위 집단(P_1, P_2, P_3, …, P_j)과 전체 집단(P) 사이에 나타난 각 동등화 결과 차이를 제곱하여 평균을 구한다. 이때 각 하위 집단의 중요도를 반영하여 가중 평균을 구한다. Dorans & Holland(2000)가 사용한 가중치 w_j는 전체 집단 P에 대한 하위 집단 P_j의 사례수의 비율이다.

분모인 $\sigma_P(Y)$는 검사 Y의 전체 집단에서의 척도점수 표준편차이다. 이렇게 표준편차로 나누어주면 RMSD값의 해석이 보다 용이해진다. 예를 들어, RMSD값이 0.1인 경우 검사 Y에서의 척도점수 표준편차의 10%임을 의미한다. RMSD값은 검사 X의 원점수 x의 개수만큼 계산되어진다. 즉 RMSD는 각 x점수에 따라 변하는 조건적인 통계치이다.

REMSD는 각 x점수에서의 RMSD(x)를 평균하여 하나의 종합적인 지표를 계산한 것이다. Dorans & Halland(2000)는 식(1)에서 계산된 RMSD값들을 각 x점수에서의 피험자 분포에 따라 가중 평균 내어 REMSD(Root Expected Mean Square Difference)라는 종합적인 지표를 계산하였다. REMSD의 계산 공식은 다음과 같이 나타낼 수 있다.

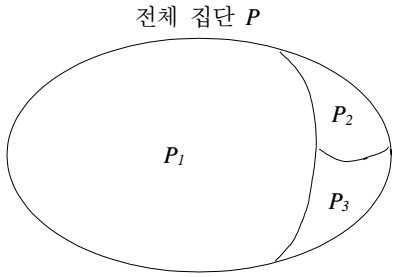
$$REMSD=\frac{\sqrt{\sum_j w_j \boldsymbol{E}[S_{p_j}(x)-S_p(x)]^2}}{\sigma_P(Y)}. \quad (2)$$

식(2)에서 E는 '평균을 계산' 한다는 의미인데 여기에서는 편차점수의 제곱 $[S_{P_j}(x)-S_P(x)]^2$을 각 x점수에 분포하는 피험자 비율을 가중치로 사용하여 가중 평균하는 것을 의미한다. 이렇게 REMSD를 계산하면 전체 점수대의 RMSD값을 평균하는 종합적인 지표를 얻을 수 있다. 이러한 종합적인 지표는 한 개의 대푯값을 제시함으로써 집단불변성 원칙을 용이하게 평가할 수 있는 장점이 있다.

즉, RMSD는 각 원점수마다 계산되는 조건적 오차 통계치이며, REMSD는 각 원점수마다 계산된 RMSD를 평균 낸 종합적 오차 통계치이다. 하지만 REMSD는 각 점수대에서의 정보를 소홀히 하는 경향이 있다. 따라서 REMSD만을 제공한다면 관심을 가져야 할 특별한 점수대에서의 현상을 관찰할 수 없다. 만약 검사 점수를 이용해 장학금 수혜자를 결정하고자 하는 경우, 낮은 점수대보다는 높은 점수대에서의 집단불변성 여부가 관심의 대상이 될 것이다. 만약 중요한 점수 범위에서 동등화 차이가 컸음에도 불구하고 다른 점수대에서 동등화 차이가 작다면, 평균을 낼 때 상쇄효과에 의해 REMSD가 작아질 수 있다. 경우에 따라서는 전체 점수대를 종합한 총체적인 집단불변성 원칙 검증보다 관심이 집중되는 중요한 점수 범위에서의 집단 불변성 원칙 검증이 더 중요할 수 있다. 가장 좋은 예로서 합격/불합격을 결정짓는 분할점수(cut-score)가 존재하는 경우, 그 분할점수 부근에서의 집단불변성 원칙 위배 여부가 최대의 관심사가 될 것이다. 따라서 이 연구에서는 RMSD와 REMSD를 모두 중요하게 다루고 있으며, 종합적 통계치인 REMSD 결과와 조건적 통계치인 RMSD 결과를 함께 확인할 것이다.

2. 사례수의 영향

Yang et al.(2003)은 두 집단 간 사례수가 현저하게 차이가 나는 경우 RMSD 및 REMSD가 실제보다 작아졌다고 보고하였다. RMSD 또는 REMSD가 사례수의 영향에 의해 작아지면 실제로는 집단불변성 원칙이 위배되었어도 그것을 발견하지 못하고 지나치게 된다. Yang(2004)은 여러 집단 간 동등화를 비교하는 경우 한 집단의 크기가 다른 집단들에 비해 두드러지게 큰 경우에 RMSD 및 REMSD의 크기가 작아지는 경향이 있는 것을 보고하였다. 그는 전체 집단 내에 사례수가 우세하게 큰 다수 집단(예를 들어, 미국 내 SAT를 치룬 백인과 기타 인종을 비교하는 경우)이 존재하면 그 다수 집단과 전체 집단 간 동등화 차이가 작은 경향이 있기 때문에 결국 작은 RMSD와 REMSD가 얻어진다고 해석하였다.



〔그림 1〕 P_1, P_2, P_3의 전체집단 P 내에서의 분할

전체 집단 P와 그에 속해있는 세 개의 하위 집단 P_1, P_2, P_3 가 있다고 가정해보자. 〔그림 1〕은 전체 집단 내 세 집단 중 P_1의 크기가 P_2 또는 P_3에 비해 현저히 우세한 경우를 보여준다. 이런 경우 P_1의 동등화 결과는 전체 집단 P의 동등화 결과와 가장 유사할 것이다. P_1은 다수 집단이므로 전체 집단 P의 특성을 가장 잘 대표하고 있는 집단이기 때문이다. 〔그림 1〕의 경우, RMSD 및 REMSD를 계산하기 위해서는 식(1)과 식(2)에서와 같이 P_1과 P 간 동등화 차의 제곱, P_2와 P 간 동등화 차의 제곱, P_3와 P 간 동등화 차의 제곱의 세 가지를 평균내야 한다. 평균을 계산할 때 각 집단(P_1, P_2, P_3)의 상대적 크기를 가중치로 사용하는데 P_1의 크기가 월등히 크므로 P_1과 P 간 동등화 차의 제곱에 해당되는 가중치가 제일 크다. 그런데 P_1과 P 간 동등화 차이는 근소하므로, 작은 값에 큰 가중치가 적용된 결과로 RMSD 및 REMSD값이 작아진다는 것이다.

〔그림 1〕에서 제시하고 있는 예와 대치되는 경우로 성별에 따른 집단을 들 수 있다. 검사를 치룬 피험자 전체집단에 속한 남·여 집단의 크기는 일반적으로 동일하거나 비슷하다. 이렇듯 집단의 크기가 균형을 이루는 경우 RMSD 또는 REMSD는 아무런 문제없이 적절히 사용될 수 있을 것이다. 요약컨대 비교하고자 하는 집단들의 사례수가 유사하지 않고 불균형할 때 RMSD 또는 REMSD값은 왜곡될 가능성이 있다.

Ⅲ. 연구 방법

1. 분석 자료

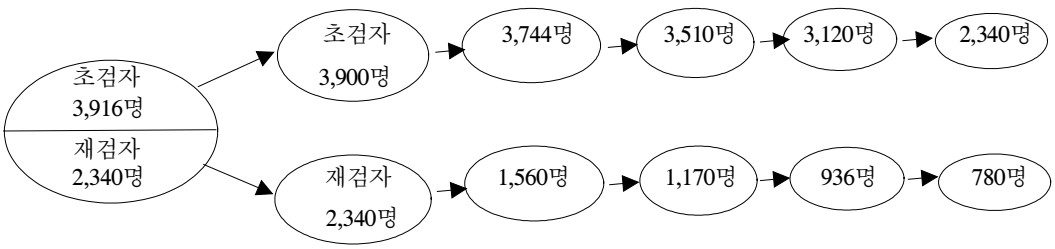
이 연구는 무선 집단 설계에 의해 최근에 치러진 표준화 학력 검사 자료를 사용하였다. 동일한 검사 시점에 치러진 검사 X, Y가 동등화에 사용되었으며 검사 X가 검사 Y로 동등화되었다. 검사 Y는 이전에 치러진 적이 있는 검사로서 이미 원점수-척도점수 전환표가 성립되어 있으므로 처음으로 실시된 검사 X를 검사 Y로 동등화한다. 검사 X를 치른 피험자 총 6,389명 중 3,916명이 초검자였고 2,473명이 재검자였으며 검사 Y를 치룬 총 3,145명 중 2,001명이 초검자였고 1,144명이 재검자였다. 앞에서 제시한 두 가지 연구문제를 탐구하기 위하여 초검자와 재검자를 구분하여 두 집단 간에 동등화 차이를 확인한다.

검사 X, Y의 기술통계량이 〈표 1〉에 제시되었다. 원점수의 평균(M)과 표준편차(SD)가 언어영역과 과학영역별로 제시되었으며, 언어영역의 원점수 범위는 0~50점, 과학영역의 원점수 범위는 0~63점이다.

〈표 1〉 검사 X, Y의 기술통계량

	N(%)	언어영역		과학영역	
		M	SD	M	SD
검사 X					
초검자	3,916(61%)	33.11	7.45	40.68	10.27
재검자	2,473(39%)	31.95	6.50	39.14	8.66
전체	6,389(100%)	32.66	7.12	40.08	9.71
검사 Y					
초검자	2,001(64%)	34.33	7.97	39.63	10.42
재검자	1,144(36%)	33.33	7.06	38.65	8.84
전체	3,145(100%)	33.96	7.66	39.28	9.89

〔그림 2〕는 다양한 사례수를 가진 초검자 집단과 재검자 집단을 추출하는 과정을 보여준다. 검사 X를 치른 전체 집단 중에서 다섯 가지 사례수를 가진 다섯 개의 초검자 집단이 추출되었다. 마찬가지로 다섯 가지 사례수를 가진 다섯 개의 재검자 집단이 추출되었다. 〔그림 2〕에서와 같이 검사 X의 초검자 3,916명 중에서 3,900명이 무선적으로 추출되었고, 선택된 3,900명 중에서 3,744명이 또 다시 무선 추출되었다. 마찬가지로 3,510, 3,120, 2,340명의 초검자 집단이 계속해서 무선 추출되었다. 또한 똑같은 방식으로 2,340, 1,560, 1,170, 936, 780명의 사례수를 가지는 다섯 가지 재검자 집단이 무선 추출되었다. 사례수를 줄여나갈 때 항상 먼저 추출된 표본에서 다음 표본을 추출하는 방식으로 최대한 표본들이 겹치도록 하여 사례수가 작아지는 것 이외에 다른 조건들은 최대한 유사하도록 설계하였다.



〔그림 2〕 검사 X에서의 표본추출

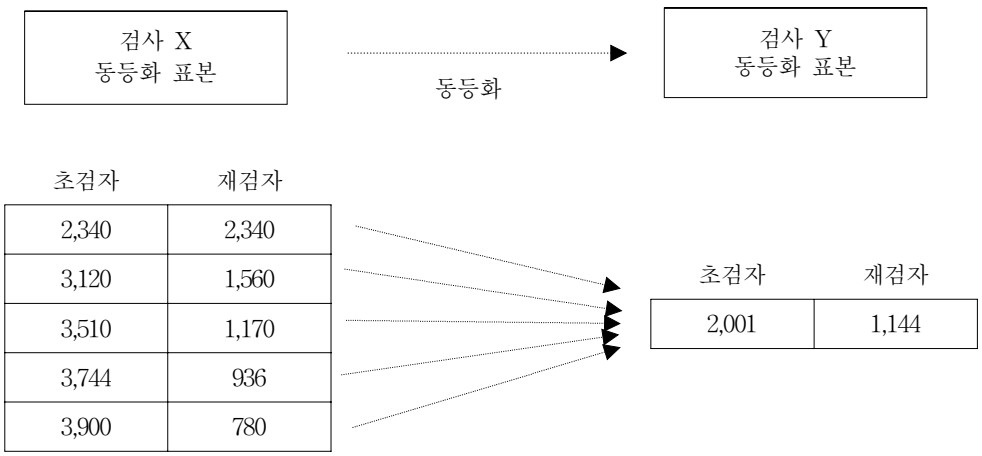
다음 단계는 〔그림 2〕에서와 같이 준비된 다섯 가지 초검자 집단과 다섯 가지 재검자 집단을 서로 짝짓는 것이다. 초검자 2,340명은 재검자 2,340명(2,340:2,340=1:1)과 결합되고, 초검자 3,120명은 재검자 1,560명(3,120:1,560=2:1)과, 초검자 3,510명은 재검자 1,170명(3,510:1,170=

3:1)과, 초검자 3,744명은 재검자 936명(3,744:936=4:1)과, 초검자 3,900명은 재검자 780명(3,900:780=5:1)과 서로 짝지어져 결합되었다. 이러한 방식으로 초검자:재검자의 비율이 1:1, 2:1, 3:1, 4:1, 5:1로 변화하도록 준비하였고 두 집단의 비율은 변화하지만 두 집단을 합한 총 사례수는 항상 4,680명으로 고정되도록 하였다. 결합된 집단(초검자 + 재검자)의 사례수를 동일하게 고정시킴으로써 전체 집단의 사례수가 미치는 영향은 배제하고 두 하위 집단 간 크기가 점점 불균형해질 때의 영향만을 탐색하고자 시도하였다. 따라서 총 사례수는 고정되어 있지만 두 집단 간 사례수가 점차적으로 불균형해질 때 RMSD 및 REMSD가 어떠한 영향을 받는지 관찰할 수 있도록 준비하였다.

2. 동등화 과정

〔그림 3〕은 본 연구에서 이용된 동등화 표본들을 나타낸다. 검사 X가 검사 Y로 동등화될 때 검사 X의 표본은 〔그림 3〕에서 나타낸 바와 같이 초검자:재검자의 비율이 1:1, 2:1, 3:1, 4:1, 5:1로 변하도록 준비되었다. 검사 Y의 표본으로서는 원래 검사 Y를 치렀던 2,001명 초검자와 1,144명 재검자가 그대로 사용되었다. 즉 검사 X 표본들은 두 집단 간 사례수 차이가 5단계로 점점 벌어지도록 조작되었고 검사 Y 표본은 검사 Y를 치룬 피험자 집단이 그대로 이용되었다.

각 검사영역(언어・과학영역)별로 초검자와 재검자 집단을 사용하여 따로 동등화를 실시하였으며 초검자와 재검자 집단의 사례수는 〔그림 3〕에서 나타난 바와 같이 변화하였다. 동등화가 실시된 후에는 집단 간(초검자 vs. 재검자) 동등화 차이를 비교하기 위한 RMSD 및 REMSD를 계산하였다.



〔그림 3〕 동등화에 이용된 표본들

동등화 방법으로는 동백분위 동등화(equipercentile equating)를 사용하였다. 다수집단과 소수집단이 존재했을 때, 소수집단의 사례수는 실제 상황에서 전체 집단의 약 20%에 그치는 경우가 빈번하다. 예를 들어, SAT와 GRE에서 백인이 아닌 인종의 비율은 약 17%였다(Pennock-Román, 1999). 따라서 소수집단의 사례수가 1천명이 넘지 않는 경우가 흔히 발생할 수 있는 문제점을 고려하여 본 연구에서는 동백분위 동등화 방법을 이용하였다. 문항반응이론을 이용한 동등화 방법을 사용하는 경우, 안정적인 문항모수 추정을 위하여 최소 1천명 이상의 표본을 대상으로 할 것을 권장하기 때문이다. 또한 문항반응이론을 이용한 동등화 방법이 동백분위 동등화 방법에 비해 검사동등화의 집단불변성 원칙을 더 잘 만족시키는 경향이 있음을 밝힌 선행 연구가 있다(Braun & Holland, 1982; van der Linden, 2000). 문항반응이론과 동백분위법을 이용한 두 가지 동등화 방법을 모두 사용하면, 집단 간 사례수 차이뿐만 아니라 동등화 방법의 차이에 따른 영향이 혼재하는 것을 방지하기 위해, 동백분위 동등화 방법 한 가지만 사용하여 집단불변성 원칙을 검증하였다.

〈표 2〉에 다섯 가지 조건에서 초검자와 재검자 집단의 평균 점수와 표준편차를 나타내었다. 1:1조건에서 초검자와 재검자 집단의 크기는 2,340명으로 동일한데, 평균점수는 재검자 집단이 더 낮고, 표준편차도 재검자 집단이 더 작다.

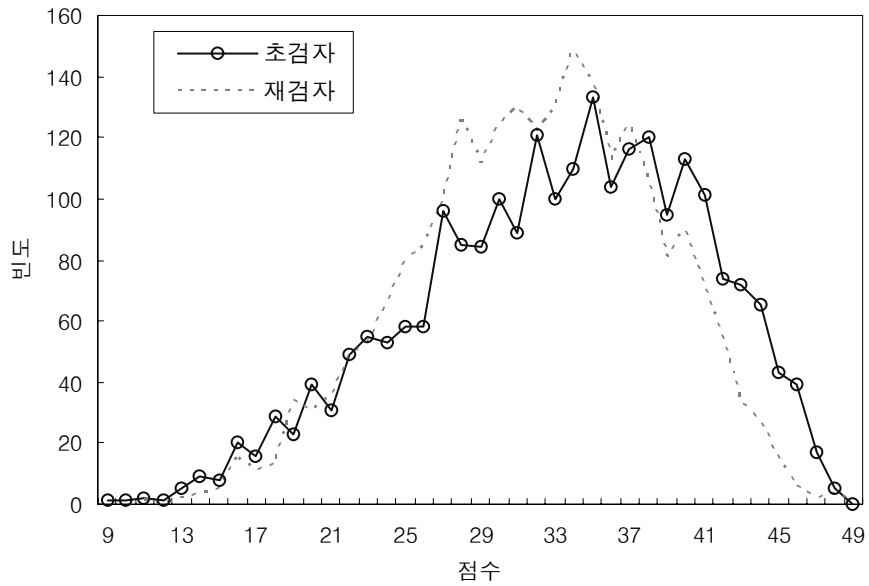
〈표 2〉 다섯 가지 동등화 표본의 기술통계량

조건	초검자					재검자				
	N	언어영역		과학영역		N	언어영역		과학영역	
		M	SD	M	SD		M	SD	M	SD
1:1	2,340	33.10	7.54	40.67	10.31	2,340	31.93	6.48	39.05	8.65
2:1	3,120	33.01	7.45	40.54	10.27	1,560	31.94	6.47	39.07	8.71
3:1	3,510	33.10	7.45	40.63	10.28	1,170	32.08	6.44	39.42	8.61
4:1	3,744	33.12	7.47	40.68	10.27	936	31.98	6.40	39.50	8.57
5:1	3,900	33.10	7.46	40.68	10.27	780	32.20	6.34	39.49	8.61

3. 가중치 부여 방법

동백분위 동등화 방법을 사용하여 영역(언어·과학영역)별로 초검자와 재검자 집단을 따로 동등화한 후 두 집단의 동등화 결과를 비교하기 위해 RMSD 및 REMSD를 계산하였다. RMSD와 REMSD를 계산하는 식(1)과 식(2)에 나타난 w_j는 평균을 계산할 때 각 집단에 부여되는 가중치이다. 집단에 가중치를 부여함으로써 그 집단이 오차 통계치에 기여하는 중요도가 결정된다.

Dorans & Holland(2000)는 각 집단의 가중치를 $w_j=N_j/N$로 결정하였다. N_j는 j번째 하위 집단의 사례수이고 N은 전체집단의 사례수이다. 항상 모든 가중치의 합은 1이 된다($\sum w_j = 1$). Dorans & Holland(2000)가 사용했던 가중치 부여 방법을 본 연구에서는 *w1*이라 칭하기로 한다. 두 번째 가중치 부여 방법으로서 원점수의 분산을 고려하여 가중치를 부여하는 방법 *w2*를 제안한다. 각 집단의 원점수 분산을 고려하여 가중치를 결정하는 방법으로서 $w2_j = V_j^{-1} / \sum V_j^{-1}$로 나타낸다. 여기에서 V_j^{-1}은 j번째 하위 집단의 원점수 분산의 역수를 취한 것이다. 이러한 두 번째 방법을 사용하면 큰 점수 분산을 가지는 집단이 작은 가중치를 부여받게 된다. 두 번째 방법을 제안하게 된 논리는 집단의 분산이 클수록 원점수의 분포가 매끄러운 곡선을 보이지 않고 불규칙한 다각형 분포를 보일 것이 예상되므로 동등화 오차가 더 클 것이기 때문이다.



〔그림 4〕 1:1조건에서 초검자와 재검자의 도수분포다각형(언어영역)

〔그림 4〕는 초검자와 재검자의 비율이 1:1조건일 때 초검자 집단과 재검자 집단의 언어영역 점수 분포를 나타낸다. 〈표 2〉에서와 같이 1:1조건에서 초검자와 재검자 집단은 각각 2,340명으로 사례수는 동일한데, 언어영역 점수의 표준편차는 초검자 집단에서 7.54, 재검자 집단에서 6.48이었다. 〔그림 4〕를 보면, 표준편차가 더 큰 초검자 집단의 도수분포다각형이 더 불규칙하였다. 특히 동백분위 동등화를 사용하는 경우에 원점수 분포가 불규칙한 다각형

모양이면 동등화 오차를 줄이기 위해 곡선화(smoothing)하는 과정을 거치기도 한다. 이 점을 고려하여 분산이 큰 집단의 동등화 결과보다 분산이 작은 집단의 동등화 결과에 오차가 적을 것이라는 예상 하에 제안한 방법이다.

세 번째 방법 *w*3는 Dorans & Holland(2000)가 사용한 *w*1을 반대로 한 것이다. 선행 연구에서 Dorans & Holland(2000)의 방법대로 가중치를 부여하였을 때 RMSD 및 REMSD가 사례수의 영향에 의해 왜곡되는 현상을 보고하였으므로 이 연구에서는 그들이 사용했던 가중치 부여 방식을 반대로 적용해본다. 기존에 이용되던 가중치 부여 방식을 단순히 반대로 적용해 보는 경우이다. 따라서 $w3_j = N_j^{-1} / \sum N_j^{-1}$ 이며 여기서 N_j^{-1}은 *j*번째 하위 집단의 사례수의 역수를 취한 것이다. 이렇게 하면 사례수가 큰 집단에게 작은 가중치가 부여된다.

네 번째 가중치 부여 방법 *w*4는 각 집단에 동일한 가중치를 주는 것으로 $w4_j = 1/n$이다. 여기서 *n*은 집단의 개수를 의미한다. 이 연구에서는 전체집단을 초검자와 재검자로 구분하므로 *n*=2이다. 네 가지 방법(*w*1, *w*2, *w*3, *w*4)에서 공통적으로 항상 가중치를 모두 합산하면 1이 된다.

위에서 밝힌 네 가지 가중치 부여 방법을 정리하면 다음과 같다.

1) $w1_j = N_j / \sum N_j$ (여기서 N_j = *j*번째 하위 집단의 사례수).

2) $w2_j = V_j^{-1} / \sum V_j^{-1}$ (여기서 V_j^{-1}= *j*번째 하위 집단의 원점수 분산의 역수).

3) $w3_j = N_j^{-1} / \sum N_j^{-1}$ (여기서 N_j^{-1}= *j*번째 하위 집단의 사례수의 역수).

4) $w4_j = 1/n$ (여기서 *n* = 하위 집단의 개수).

Ⅳ. 연구 결과

이 연구에서는 3가지 요인을 고려한 분석을 실시하였다. 첫째는 다섯 단계(1:1, 2:1, 3:1, 4:1, 5:1)로 변하는 집단 간 사례수 차이이고, 둘째는 RMSD 및 REMSD 계산 과정에서 사용되는 네 가지 가중치 부여 방법(*w*1-*w*4)이며, 셋째는 두 가지 검사 영역(언어・과학)이다. 따라서 총 40가지(5×4×2)의 조건에 따라 RMSD 및 REMSD가 계산되었다.

1. REMSD (종합적 통계치)

〈표 3〉 네 가지 가중치를 이용하여 계산된 REMSD (언어영역)

초검자:재검자	*w*1	*w*2	*w*3	*w*4
1:1	.0607	.0582	.0607	.0607
2:1	.0400	.0462	.0485	.0444
3:1	.0556	.0723	.0800	.0689
4:1	.0655	.0938	.1071	.0888
5:1	.0716	.0989	.1123	.0942

REMSD는 모든 점수대에서의 RMSD값들을 평균한 값으로 한 개의 종합적 지수로 집단불변성 원칙을 평가할 수 있기 때문에 편리하다. 언어영역을 위한 REMSD값들이 〈표 3〉에, 과학영역을 위한 REMSD값들이 〈표 4〉에 제시되었다. 1:1 조건은 초검자와 재검자 집단의 사례수가 동일한 경우이므로 다른 조건에서 계산된 통계치와 비교할 수 있는 기준(criteria)이 될 수 있다. 두 집단 간에 크기가 균형을 이루는 경우(1:1 조건)의 REMSD를 참값이라 가정하고 이와 비교해서 집단 간 크기가 2:1, 3:1, 4:1, 5:1로 벌어질 때 어떠한 경향을 보이는지 관찰하였다.

〈표 3〉에서 2,340명의 초검자와 2,340명의 재검자를 이용하여 동등화를 실시한 1:1 조건의 경우, *w*1, *w*3, *w*4의 세 가지 가중치는 일치하는 REMSD값을 보였다. 두 집단의 크기가 1:1로 동일한 경우, 세 방법은 결국 동일한 가중치를 사용하게 되기 때문이다. 1:1 조건의 경우, *w*2 방법도 다른 세 가지 방법(*w*1, *w*3, *w*4)을 사용했을 경우와 매우 가까운 REMSD값을 산출하였다. 3,120명의 초검자와 1,560명의 재검자를 이용하여 동등화를 실시한 2:1 조건은 1:1 조건에 비교하여 작아진 REMSD값을 보였다. 이러한 경향은 네 가지 가중치 부여 방식(*w*1-*w*4)에서 공통적이었다. 그러나 3,510명의 초검자와 1,170명의 재검자가 결합된 3:1 조건에서는 네 가지 가중치 부여 방식에서 모두 2:1 조건에 비하여 다시 증가한 REMSD값을 보였다. 즉 REMSD값은 2:1 조건에서는 줄어들었으나 3:1 조건에서는 다시 증가하기 시작하여 5:1 조건까지 계속해서 커졌다. 따라서 5:1 조건에서의 REMSD는 1:1 조건에서의 REMSD 보다 큰 값을 보였다. 또한 *w*1, *w*3, *w*4가 모두 동일한 REMSD값을 보였던 1:1 조건을 제외하고는 2:1부터 5:1 조건까지 REMSD는 *w*3 > *w*2 > *w*4 > *w*1의 순서로 큰 값을 보였다.

〈표 4〉에 나타난 바와 같이 과학영역도 언어영역과 비슷한 패턴을 보여준다. 하지만 REMSD가 다시 커지는 시점은 언어영역의 경우와 차이가 있었다. 〈표 4〉에서 REMSD는 3:1 조건까지 줄어들지만 4:1 조건부터 커지기 시작하여 5:1 조건까지 증가하였다. 언어영역에서와 마찬가지로 1:1 조건만 제외하고는 *w*3 > *w*2 > *w*4 > *w*1의 순서로 큰 REMSD값을 산출하였다.

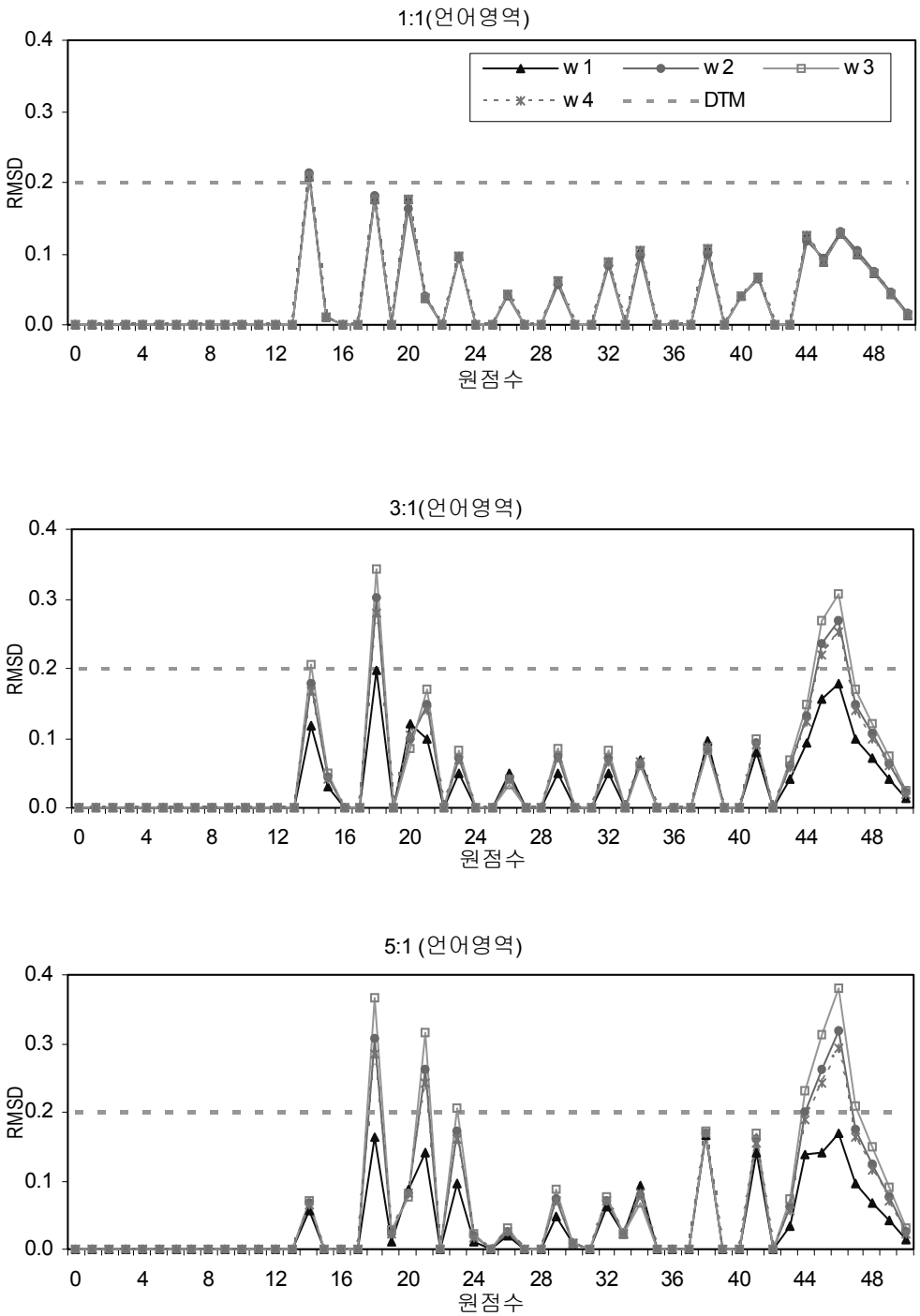
〈표 4〉 네 가지 가중치를 이용하여 계산된 REMSD (과학영역)

초검자:재검자	$w1$	$w2$	$w3$	$w4$
1:1	.0908	.0872	.0908	.0908
2:1	.0631	.0755	.0793	.0717
3:1	.0575	.0718	.0778	.0684
4:1	.0652	.0804	.0876	.0772
5:1	.0764	.0926	.1009	.0895

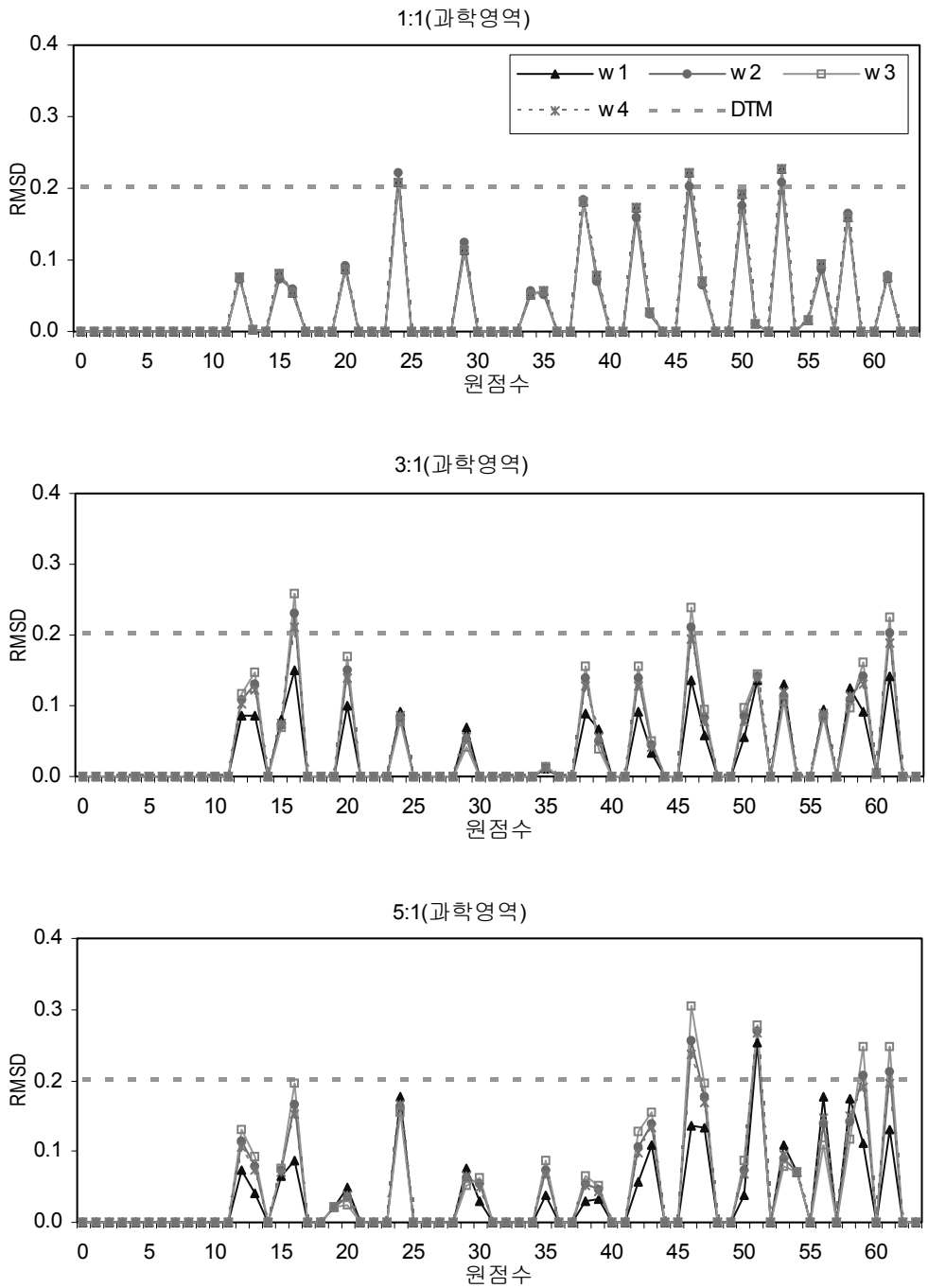
2. RMSD (조건적 통계치)

〔그림 5〕에 포함된 세 개의 그래프는 언어영역에서 초검자:재검자 비율이 각각 1:1, 3:1, 5:1일 때 RMSD값을 보여준다. 각 그래프에는 RMSD의 크기를 평가하기 위해 DTM (Difference That Matters)을 함께 표시하였다. Dorans & Feigenbaum(1994)이 동등화 차이를 비교할 때 처음으로 DTM을 사용하였다. DTM은 척도점수 단위의 반을 표준화한 값으로 정의할 수 있다. 본 연구에서 사용된 검사의 척도점수 단위는 1점이므로 그 반인 0.5를 표준편차로 나누면 DTM을 얻을 수 있다. DTM을 기준으로 RMSD의 크기를 판단할 수 있는 원리는 두 개의 동등화 결과가 척도점수 단위의 반을 넘는 차이를 보인다면 반올림 과정을 거치고 나면 두 동등화 결과는 서로 다른 척도점수를 갖게 되기 때문이다. RMSD가 DTM보다 작으면 동등화 차이가 염려할 만큼 크지 않다고 판단할 수 있으므로 집단불변성 원칙을 만족한다는 결론을 내릴 수 있다. 언어영역에서 척도점수 표준편차는 2.49였으므로 DTM은 0.5/2.49= 0.2008이다. 〈표 3〉과 〈표 4〉에서와 마찬가지로 1:1 조건에서의 RMSD값이 비교의 기준으로 사용될 수 있다.

〔그림 5〕에서 첫 번째 그래프는 초검자:재검자 비율이 1:1인 조건에서 초검자를 이용한 동등화 결과와 재검자를 이용한 동등화 결과를 비교하여 계산한 RMSD값들을 보여주고 있다. REMSD와 마찬가지로 $w1$, $w3$, $w4$의 세 가지 가중치 부여 방법은 전체 점수대에 걸쳐 동일한 RMSD값을 산출하였다. $w2$도 $w1$, $w3$, $w4$와 매우 유사한 RMSD를 보이므로 네 가지 가중치는 거의 겹치는 그래프를 보이고 있다. 두 집단의 크기가 비슷할 때는 어떠한 가중치 부여 방법을 사용하여 RMSD를 계산하여도 큰 차이가 없음을 알 수 있다.



〔그림 5〕 언어영역에서 초검자:재검자 비율이 1:1, 3:1, 5:1일 때 RMSD



〔그림 6〕 과학영역에서 초검자:재검자 비율이 1:1, 3:1, 5:1일 때 RMSD

〔그림 5〕에서 1:1 조건에서는 검사 X의 원점수가 14점일 때 네 가지 가중치 부여 방법(*w*1-*w*4) 모두 DTM 보다 큰 RMSD값을 보였다. 본 연구의 자료로 쓰인 시험에서 합격/불합격을 결정하는 분할점수는 학교마다 다르지만 전체 점수대를 하위/상위 점수대로 반분한다면 분할점수는 항상 상위 점수대에 속하기 때문에 하위 점수대에 위치하는 14점은 크게 관심의 대상이 되는 점수는 아니다. 따라서 집단불변성 원칙은 이 시험의 용도를 고려해볼 때 만족되었다고 볼 수 있다. 두 번째 그래프는 3:1 조건에서 계산된 RMSD값을 보여주는데, 네 가지 가중치 부여 방법(*w*1-*w*4)으로부터 계산된 RMSD값들은 이제 서로 현저하게 차이가 나는 것을 관찰할 수 있다. *w*3으로부터 계산된 RMSD는 원점수가 14, 18, 45, 46점일 때 DTM 보다 큰 값을 보였고, *w*2와 *w*4로부터 계산된 RMSD는 원점수가 18, 45, 46점일 때 DTM을 초과하는 값을 보였으며 *w*1은 전체 점수대에서 DTM보다 작은 RMSD를 산출하였다. 두 개의 원점수 45, 46점은 상위 점수대에 속하므로 이 두 점수에서의 집단불변성 위배 여부는 관심의 대상이 될 수 있다. 세 번째 그래프를 보면 5:1 조건에서의 RMSD값들은 가중치 부여 방법에 따라 차이가 더욱 두드러졌다. *w*3은 7개의 원점수(20, 23, 25, 46-49점)에서 DTM 보다 큰 RMSD를 산출하였고, *w*2와 *w*4는 4개의 원점수(20, 23, 47, 48점)에서 DTM보다 큰 RMSD를 산출하였으며, *w*1은 3:1 조건에서와 마찬가지로 전체 점수대에서 DTM보다 작은 RMSD를 산출하였다. 따라서 언어영역의 경우 1:1 조건에서는 네 가지 가중치 부여 방법이 모두 집단불변성 원칙을 만족시켰고, *w*1의 경우 두 하위 집단 간 사례수 차이가 3:1 또는 5:1로 벌어져도 집단불변성 원칙을 만족시키는 것으로 나타났다. *w*1을 제외한 나머지 가중치 부여방식은 3:1과 5:1 조건에서는 집단불변성 원칙을 만족시키지 못하였다.

〔그림 6〕은 과학영역에서의 RMSD값을 보여주는 세 개의 그래프를 포함하고 있다. 과학영역에서 척도점수 표준편차는 2.47이었으므로 DTM은 0.5/2.47=0.2024이다. 언어영역에서와 마찬가지로 1:1에서 5:1 조건으로 갈수록 네 가지 가중치 부여 방법에 따른 RMSD값의 차이가 벌어짐을 볼 수 있다. 또한 첫 번째 그래프를 살펴보면, 1:1조건에서는 언어영역에서와 마찬가지로 네 가지 가중치 부여 방법을 나타내는 라인들이 거의 겹쳐있음을 볼 수 있다. 1:1 조건에서는 네 가지 가중치 부여 방법 모두 공통적으로 26, 46, 53점에서 DTM보다 큰 RMSD값을 보였다. 46점과 53점은 상위 점수대에 속하므로 관심의 대상이 되는 점수이다. 과학영역은 언어영역보다 관심 점수대에서 더 자주 집단불변성 원칙을 위배했음을 알 수 있다. 두 번째 그래프(3:1 조건)에서 *w*2와 *w*3은 18, 46, 61점에서 DTM보다 큰 RMSD값을 보였고, *w*4는 18점에서 DTM보다 큰 RMSD값을 보였으나 18점은 분할 점수가 되기에는 낮은 점수이므로 관심의 대상은 아니다. *w*1은 3:1 조건에서 항상 DTM보다 낮은 RMSD값을 보였다. 세 번째 그래프(5:1 조건)에서 *w*2와 *w*3은 46, 51, 59, 61점에서 DTM을 초과하는 RMSD를 산출하였고, *w*1과 *w*4는 51점에서 DTM을 초과하는 RMSD를 보였다. 과학영역에서의 RMSD 결과를 요약하자면, 1:1 조건에서는 네 가지 가중치 부여 방식 모두가 집단불변성 원칙을 만족시키

지 못하였고 3:1 조건에서는 *w*1과 *w*4를 사용하였을 때 집단불변성 원칙을 만족시켰으며, 5:1 조건에서는 다시 모든 가중치 부여 방식이 집단불변성 원칙을 만족시키지 못했다.

Ⅴ. 결론 및 논의

미국에서는 검사동등화를 도입하는 검사의 종류가 꾸준히 늘어나 현재에 이르러서는 대학 또는 대학원 입학시험, 각종 면허와 자격증 검사 등 중요도가 높은 시험일수록 검사동등화를 거쳐 점수의 공정성을 유지하도록 노력하고 있다. 또한 검사 동등화와는 또 다른 목적으로, 두 가지 점수를 연계해야 하는 상황이 증가하고 있다. 예를 들어 비슷한 목적의 두 가지 검사 점수를 비교하기 위해 (예: ACT와 SAT점수의 비교) 양 점수를 일치화(concordance) 하거나, 연도별 학업 성장을 탐색하기 위해 점수의 종단적 비교를 위한 수직적 점수 연계(vertical scaling)를 해야 할 필요성이 증가하고 있다. 동등화 및 그 밖의 점수 연계 기법의 중요성과 사용 범위가 넓어짐에 따라 그 얻어진 결과의 양호성을 평가해야할 필요성도 함께 증가하였다.

한국에서도 국가수준 학업성취도 평가에서 초6, 중3, 고1 학생들의 연도별 변화 추이를 살펴보고자 동등화가 도입되어 시행되고 있다. 한국의 수능이나 국가수준의 임용고시 등에서 동등화가 도입되는 것은 문항 보안 유지와 관련된 어려움에 의해 쉬운 일은 아니지만, 그밖에 각종 면허 또는 자격증 시험, 특히 컴퓨터화 검사에서 동등화가 도입될 가능성은 충분히 있다고 판단된다. 집단불변성 원칙을 검증하는 것은 동등화 및 그와 유사한 점수 연계 결과의 양호성을 평가할 수 있는 수단이며 주로 REMSD를 계산하여 집단 불변성 위배 여부를 평가하여 왔다.

하지만 선행연구들에서 RMSD 및 REMSD가 집단의 크기에 영향을 받는 경향이 있다는 제시를 한 바가 있다. 그렇다면 집단 간 사례수가 크게 차이 나는 경우, RMSD 및 REMSD를 계산하여 집단불변성 원칙을 평가하는데 문제가 있다는 것을 암시한다. 선행연구에 따르면, 비교하려는 집단들의 크기가 불균형할 때 RMSD 및 REMSD는 작아지는 경향을 보인다고 하였다. 집단의 사례수가 REMSD에 어떠한 영향을 미치는지 정확히 파악하는 것은 집단불변성 원칙의 검증 결과에 대해서 주의를 기울여 해석할 수 있도록 하는 데 도움이 될 것이다.

본 연구에서는 비교하려는 두 집단 간에 사례수가 1:1, 2:1, 3:1, 4:1, 5:1로 변화할 때 RMSD 및 REMSD가 어떠한 영향을 받는지 조사하였다. 집단불변성 연구에서 시뮬레이션 연구가 어려운 점은 성별, 인종이나 재검여부 등으로 구분되는 집단의 특성을 고려하여 데이터를 생성하기가 어렵기 때문이다. 실제로 집단불변성에 관련된 선행연구에서 시뮬레이션 연구를 찾아볼 수 없다. 따라서 이 연구에서는 사례수가 많은 실제 데이터에서 재검자와 초

검자 집단을 구분한 유층표집(stratified sampling)을 실시하여 여러 가지 사례수를 조작하였다. 집단의 크기가 1:1인 조건에서 계산된 RMSD 및 REMSD를 참값으로 취하고 비교의 기준으로 삼았다. 두 집단 간 크기가 동일할 때 계산된 RMSD 및 REMSD에 비교하여 2:1, 3:1, 4:1, 5:1로 집단 간 크기 차이가 점점 커질수록 RMSD 및 REMSD가 어떠한 영향을 받는지 관찰하였다. 또한 RMSD 및 REMSD를 계산할 때 사용되는 가중치를 네 가지 방법으로 달리 적용하여 어떠한 가중치 부여 방법이 집단의 사례수로부터 가장 영향을 덜 받는지 관찰하였다. 결과적으로 이 연구에서는 다섯 단계(1:1, 2:1, 3:1, 4:1, 5:1)로 변하는 집단 간 사례수 차이를 고려하였고, RMSD 및 REMSD 계산 과정에서 사용되는 네 가지 가중치 부여 방법($w1$-$w4$)의 적용을 시도하였으며, 두 가지 검사 영역(언어・과학)을 대상으로 동백분위 동등화를 실시하였다. 따라서 총 40가지(5×4×2)의 조건에 따라 RMSD 및 REMSD가 계산되었다.

〈표 3〉과 〈표 4〉에서 제시한 종합적 통계치인 REMSD를 관찰한 결과, 언어영역에서는 집단 간 크기가 2:1로 차이나는 경우에 REMSD가 작아졌으나 3:1 조건부터는 다시 커지는 경향을 보였으며 과학영역에서는 3:1 조건까지는 작아졌으나 4:1 조건부터 다시 커지기 시작하였다. 선행연구에서는 집단 간 크기 차이가 클수록 REMSD가 작아지는 경향이 있다고 하였으므로 1:1 조건에서 5:1 조건으로 갈수록 REMSD가 점점 작아질 것을 예상하였다. 하지만 본 연구의 결과는 계속적으로 작아지는 것이 아니라 집단 간 차이가 언어영역에서는 2:1, 과학영역에서는 3:1까지 REMSD가 작아지는 경향을 보이다가 그 이상으로 차이가 벌어지면 REMSD가 다시 커지기 시작하는 경향을 보였다. 이렇게 REMSD가 작아지다가 다시 커지는 시점이 언어영역과 과학영역에 따라 차이가 있는 것은 사례수의 영향이 교과목과 상호작용하는 현상이 있음을 알 수 있다. 〈표 3〉과 〈표 4〉를 살펴보면 기준이 되는 1:1 조건에서 과학영역이 언어영역에 비해 큰 REMSD를 보이고 있으므로 이 연구에 쓰인 자료의 경우, 과학영역이 언어영역에 비해 원래 집단불변성 원칙 위배 정도가 높다고 볼 수 있다. 또한 〔그림 5〕와 〔그림 6〕에서도 언어영역의 경우 1:1 조건에서 모든 가중치 부여 방식이 집단불변성 원칙을 만족시켰으나, 과학영역의 경우 1:1 조건에서 네 가지 가중치 부여 방식 모두가 집단불변성 원칙을 만족시키지 못하였다. Kolen(2004)도 검사동등화의 집단불변성 원칙 위배 여부는 과목에 따라 다를 수 있다고 하였다. 검사 동등화의 집단불변성 위배 정도가 교과목에 따라 달라지듯이 사례수의 영향도 교과목에 따라 달라졌다. 이러한 교과목에 따라 달라지는 집단불변성 위배 여부는 집단의 사례수가 REMSD에 미치는 영향과 상호작용을 했을 가능성이 있음을 알려준다. 따라서 언어・과학영역 이외에 다른 교과에서의 경향도 연구해볼 필요가 있다.

그렇다면 왜 예상했던 대로 집단 간 크기 차이가 커질수록 REMSD가 작아지지 않았을까? Yang(2004)의 연구에서는 전체 집단이 다수 집단과 소수 집단으로 나뉠 때 언어영역에서의 동등화 결과를 대상으로 REMSD를 계산하였다. Yang의 연구에서 소수 집단의 최소 크기는 1,190명이었다. 본 연구에서 소수 집단의 최소 사례수는 780명이었다. 즉 선행연구에서는 소

수 집단이 1000명 보다 작은 조건에서 REMSD의 경향성을 확인하지 않았다. 본 연구에서는 3:1조건에서 작은 집단은 1,170명, 4:1조건에서 작은 집단은 936명으로 구성되었고, 5:1조건에서 작은 집단은 780명의 사례수를 가졌다. 분석 결과, 언어영역에서는 3:1조건부터 REMSD가 커지기 시작하였고, 과학영역에서는 4:1조건부터 REMSD가 커지기 시작하였다. 결국, 이 결과로부터 한쪽 집단의 사례수가 너무 작아지면 REMSD가 커지고, 그 시작점은 과목 간에 차이가 있다는 결론이 된다.

너무 작은 사례수는 안정적이지 못한 동등화 결과를 산출할 수 있다. 이러한 큰 동등화 오차가 전체 집단과 780명 집단 간에 동등화 차이를 크게 할 수 있고 작은 가중치가 주어짐에도 불구하고 상당히 큰 REMSD를 산출하는데 기여할 수 있다. 따라서 집단의 크기가 매우 작은 경우에 동등화를 실시하는 경우 동등화 오차가 커지므로 동등화 오차를 고려하지 않는다면 집단불변성 원칙을 정확히 검증할 수 없다는 것을 알 수 있다. 사례수가 아주 작은 집단을 포함하여 집단불변성 원칙을 검증하고자 하는 경우 REMSD와 같은 통계치를 계산하는 것보다 동등화 오차를 판단의 기준으로 이용하는 방법을 고안해야 할 것이다. 본 연구에서는 3:1 또는 4:1 조건부터 REMSD가 커지기 시작하였으므로 약 1000명 이하의 사례수를 가진 집단의 동등화에서 동등화 오차를 심각하게 고려해야함을 시사한다. 동백분위 동등화의 동등화 오차를 구하기 위해서는 부트스트랩 기법을 이용하거나 또는 Lord(1982)가 제안한 공식으로 계산할 수 있다.

종합적 통계치인 REMSD는 한 개의 종합적 지수로 집단불변성 원칙을 평가할 수 있기 때문에 유용하긴 하지만 특별한 관심을 가져야 할 점수대에서의 집단불변성 원칙 위배 여부를 관찰할 수 없다는 단점이 있다. 따라서 〔그림 5〕와 〔그림 6〕에 각 점수에 따른 조건적 지수인 RMSD가 제시되었다. RMSD를 살펴보면 1:1 조건에서 5:1 조건으로 갈수록 집단불변성 원칙을 더 위배하게 되는 경향이 있음을 관찰할 수 있다. 하지만 이것은 〈표 3〉과 〈표 4〉에 나타난 종합적 통계치인 REMSD의 경향과는 다르다. 종합적 통계치는 1:1 조건에서 5:1 조건으로 갈수록 계속 커지지 않고 작아지다 커졌다. 언어영역에서는 3:1 조건이 1:1 조건보다 더 작은 REMSD를 보였었고 과학영역에서는 5:1조건이 1:1 조건보다 더 작은 REMSD를 보였다. 따라서 종합적 통계치인 REMSD 또는 조건적 통계치인 RMSD 중 어느 통계치를 사용하여 집단불변성 원칙을 판단하는가에 따라 집단의 사례수에 의한 영향이 달라지는 것을 알 수 있다. 많은 집단불변성 원칙 관련 연구에서 종합적 통계치인 REMSD만을 계산하여 판단하는 경향이 있으나 조건적 통계치도 함께 검토해야 할 것이다.

연구결과, 비교하려는 집단들의 크기가 비슷한 경우(1:1 조건)에는 네 가지 가중치 부여 방법($w1$-$w4$) 중 어느 것을 사용하여도 무방함을 알 수 있었다. 하지만 1:1 조건을 제외하고는 $w1$이 집단 간 크기 차이가 벌어짐에 따라 가장 안정적인 REMSD를 산출하였고, $w4$가 두 번째로 안정적이었다. 또한 $w4$는 $w1$과 가장 가까운 오차 통계치를 산출하였다. $w4$는 각 집단에 동일한 가중치를 부여하는 방식으로 $w1$ 대신에 편리하게 쓰일 수 있을 것이다.

참 고 문 헌

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating.* New York: Academic.

Dorans, N. J. (2004). Using subpopulationinvariance to assess test score equity. *Journal of Educational Measurement*, *41*(1), 43-68.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In Lawrence, I. M., Dorans, N. J., Feigenbaum, M. D., Feryok, N. J., Schmitt, A. P., & Wright, N. K, *Technical Issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281-306.

Dorans, N. J., Holland, P. W., & Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender group for three Advanced Placement Program examinations. In Dorans, N. J. (Ed.). *Population invariance of score linking: theory and applications to Advanced Placement Program examinations* (ETS RR-03-27). Princeton, NJ: Educational Testing Service.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, *41*(1), 3-14.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, *7*, 165-174.

Pennock-Román, M. (1999). *English Proficiency and Differencies Among Racial and Ethnic Groups in Mean SAT and GRE Scores: A Longitudinal Analysis.* (ETS GBR-86-09cP). Princeton, NJ: Educational Testing Service.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*, *65*(4), 437-456.

von Davier, A. A., & Han, N. (2004). *Population invariance and linear equating non-equivalent groups design.* Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: their relationship to population invariance. *Journal of Educational Measurement*, *41*(1), 15-32.

von Davier, A. A., & Willson, C. (2004). *Population invariance of IRT equating for Advanced Placement (AP) Program Exams*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Yang, W.-L. (2004). Sensitivity of linking between AP multiple-choice scores and composite scores to geographical region: an illustration of checking for population invariance. *Journal of Educational Measurement*, *41*(1), 33-41.

Yang, W.-L., Dorans, N. J., & Tateneni, K. (2003). Effect of sample selection on Advanced Placement multiple-choice score to composite score linking. In Dorans, N. J. (Ed.). *Population invariance of score linking: theory and applications to Advanced Placement Program examinations* (RR-03-07). Princeton, NJ: Educational Testing Service.

Yi, Q., Harris, D. J., & Gao, X. (2004). *Invariance of IRT equating across different sub-populations taking a science achievement test*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

Zeng, L., Kolen, M. J., & Hanson, B. A. (1995). *Random Groups Equating Program (RAGE), version 2.0*. Iowa City, IA: ACT.

• *논문 접수 : 2009년 5월 1일 / 수정본 접수 : 2009년 6월 3일 / 게재 승인 : 2009년 6월 17일*

ABSTRACT

The Effect of Subgroup Size Imbalance on Population Invariance of Equating

HeeKyoung Kim
(Associate Research Fellow, Korea Institute for Curriculum and Evaluation)

Population invariance studies usually use error statistics such as REMSD to assess the magnitude of equating differences. However, previous research indicated that the use of error statistics for comparing equating functions between subgroups which differ considerably in sample size tended to produce small values of statistics and led to misinterpretation of whether the equating was population invariant. This study aims to examine how error statistics are sensitive to different sample sizes of subgroups. Also, alternative weighting schemes are implemented in computing error statistics to identify the weighting scheme that is the least sensitive to subgroup size. This study found that error statistics could increase or decrease solely due to different sample size. The weighting scheme that assigns the subgroup weights proportional to sample sizes of the subgroups produced the smallest changes in error statistics as the subgroup sizes varied.

Key words : Test equating, Population invariance property, Equipercentile equating, REMSD, RMSD