교육과정평가연구
The Journal of Curriculum & Evaluation
2006, Vol. 9, No. 1, pp. 69～88

대학수학능력시험의 안정적 등급 산출을 위한 요건 탐색

양 길 석(한국교육과정평가원 연구원)
민 경 석(한국교육과정평가원 연구원)
손 원 숙(한국교육과정평가원 연구원)
이 명 애(한국교육과정평가원 연구원)

≪ 요 약 ≫

교육인적자원부가 발표한 2008 대입개선안에 따르면, 2008학년도 대학수학능력시험(수능)부터 다른 점수 형태는 제공되지 않고, 영역/과목별 등급만 제공하게 된다. 현행 수능에서는 등급 이외에 표준점수, 백분위가 함께 제공되어, 상호보완적 역할을 하고 있다. 그러나 2008학년도부터 수능의 점수 체제가 등급제로 전환됨에 따라, 각 영역/과목별 뿐 아니라 영역/과목 간 균형있는 등급 점수의 산출이 무엇보다도 중요한 문제로 대두되었다. 본 연구는 2008학년도 등급제 실시에 앞서, 영역/과목별 안정적 등급 산출을 위한 조건을 탐색하기 위한 것이다. 연구의 목적을 달성하기 위하여, 첫째, 원점수, 선형 표준점수, 정규화 표준점수 등 세 가지의 검사 점수 유형에 따른 등급 산출 결과를 비교하였다. 둘째, 문항 난이도에 따른 등급 산출 결과를 비교하였다. 분석 결과, 원점수, 정수화된 선형 표준점수 및 정규화 표준점수 분포를 이용하여 등급 비율을 산출했을 때, 상대적으로 점수의 가지 수가 많은 언어, 수리, 외국어(영어) 영역의 경우는 검사 점수 유형에 따른 등급별 비율에 큰 차이가 없는 것으로 나타났다. 반면 탐구 및 제2외국어/한문 영역의 선택과목과 같이 점수 가지 수가 적은 시험에서는 가능한 모든 점수 가지 수를 살린 분포, 즉 원점수 분포를 이용하여 등급 점수를 산출하는 것이 가장 안정적인 것으로 나타났다. 문항 난이도와 등급 비율간의 관계에서는 쉬운 문항, 중간 수준 문항, 어려운 문항 등 다양한 수준의 난이도를 갖는 검사에서 등급 비율이 안정적으로 산출되었다. 결론적으로 현행 등급 체제에서 안정적 등급 산출을 위해서는 가능한 한 원점수 분포를 그대로 사용하고, 문항 난이도의 수준이 고르게 나올 수 있도록 출제 과정을 강화하는 것이 주요 요건이라 할 수 있다. 그러나 특히 문항 수가 적은 선택과목의 경우, 문항의 난이도가 고르게 분포되도록 출제하는 것은 현실적으로 쉬운 문제는 아니다. 따라서 안정적 등급 산출을 위해서는 중·장기적으로 선택과목의 통합·조정을 통해 시험과목의 수를 축소하고 문항 수를 확대하는 방향으로 시험 체제를 개선하는 노력이 이루어져야 할 것이다.

주제어 : 등급제, 검사 점수 유형, 문항 난이도

Ⅰ. 서론

『대학수학능력시험 10년사』(한국교육과정평가원, 2005)를 살펴보면, 1994학년도부터 도입·시행된 대학수학능력시험(이하 수능)은 교육과정 개정, 시행·관리 요건의 개선, 시대·사회적 요구 등에 의하여 출제 영역과 과목, 문항 수 등 시험 체제의 변화가 몇 차례 있었고, 그에 따라 필요한 경우 점수 체제의 변화도 이루어져왔다.

1994~1998학년도는 '원점수 시기'라 할 수 있는데, 보고 점수로는 영역별 원점수와 백분위, 원점수 총점과 백분위가 제공되었다. 1999~2001학년도는 '표준점수 도입·적용 시기'라 할 수 있다. 1999학년도 수능은 6차 교육과정에 의하여 교육을 받은 수험생이 처음으로 수능을 응시하게 된 시기로, 탐구 영역에 선택과목이 도입되었다. 이에 선택과목 간 난이도 조정과 점수 비교의 합리성을 높이기 위하여 공통문항 점수를 이용한 점수 조정 절차와 표준점수 체제가 도입되었던 것이다. 이 시기에는 영역별 원점수와 백분위, 원점수 총점과 백분위, 영역별 표준점수와 변환표준점수[1], 변환표준점수 총점과 백분위가 제공되었다. 단, 2001학년도부터 실시된 제2외국어 영역의 경우에는 변환표준점수가 제공되지 않았다. 2002~2004학년도는 '총점·소수점 표기 폐지 및 등급제 도입·적용 시기'라 할 수 있는데, 이 시기에는 이전까지 제공되었던 원점수 총점과 변환표준점수의 총점을 없애는 대신 변환표준점수에 따른 종합 등급만을 제공하였다. 그리고 소수 둘째자리까지 제공되었던 점수를 소수 첫째 자리에서 반올림하여 정수로 제공하였으며, 스테나인(Stanine) 척도에 기반한 9등급제가 도입·적용되었다. 즉, 이 시기에 제공된 점수 형태는 영역별 원점수와 백분위, 영역별 표준점수, 영역별 변환표준점수와 백분위, 영역별 등급, 그리고 5개 영역 종합등급이었다. 2005학년도부터 현재까지는 '영역/과목별 표준점수 정착 시기'라 할 수 있다. 학교교육에 새롭게 적용된 제7차 교육과정에 따라 2005학년도 이후 수능은 고등학교의 심화선택과목 중심으로 시험 과목 수가 확대되고, 수험생들은 영역이나 과목을 자신의 선택에 따라 자유롭게 응시할 수 있게 되었다. 이에 상이한 시험 과목 간 점수 비교의 합리성을 높이기 위해 원점수 형태의 점수 보고 체제를 폐기하고 표준점수와 그 백분위 및 등급 점수를 제공하고 있다(남명호 외, 2002).

그런데 2008학년도 수능부터 또 한번의 점수 체제상의 변화를 겪어야 한다. 교육인적자원부는 2004년 10월 28일 『학교교육 정상화를 위한 2008학년도 이후 대학입학제도 개선안』(이하 2008 대입 개선안, 교육인적자원부, 2004)을 발표하였다. 이 개선안은 대입전형에서 학

1) 변환표준점수란 영역별 원점수의 배점 비율(언어 1.2, 수리·탐구 I 0.8, 수리·탐구 II 1.2, 외국어 0.8)을 반영하여 각 영역별 표준점수의 최고점의 합이 원점수 총점인 400점이 되도록 표준점수를 변환한 점수이다.

교교육의 과정과 결과를 중시하고 대학 자율화·특성화와 연계한 전형을 다양화하며, 선발 경쟁에서 입학 후 교육 경쟁으로 전환한다는 기본 방향을 설정하고, 학교생활기록부 반영 비중 확대, 수능 제도 개선, 대입전형의 특성화·전문화 강화, 사회통합 유도 전형 활성화 등 네 가지 핵심 과제를 제시하였다. 그 중 수능 제도 개선 과제의 핵심 내용은 선택 대상 과목 수의 축소 방안 검토, 수능과 고교 교육과정과의 연계 강화, 현행 점수 체제에서 등급만 제공하는 점수 체제로의 전환, 문제은행식 출제 방식으로의 전환, 복수 시행 방안 검토 등이다. 이 내용 중, 수험생을 비롯한 입학관계자에게 가시적으로 인식되는 변화 내용은 등급만 제공하는 점수 체제라 할 수 있다.

2008 대입 개선안에 따르면, 수능 등급제 도입 배경은 대입전형에서 수능의 영향력을 완화하고 학생부 중심의 대입전형을 유도함으로써 학교교육을 정상화한다는 것이다. 수능의 9등급제는 2002학년도부터 도입되어 다른 점수들과 함께 제공되어 왔던 것으로, 완전히 새로운 것은 아니다. 단지 2008학년도 수능부터는 다른 점수 형태를 제공하지 않고, 영역/과목별로 등급만 제공한다는 점에서 제도 변화가 된 것이다.

이처럼 등급제로 완전 전환될 경우에, 무엇보다도 가장 중요한 것은 매 시험에서 9개의 등급이 안정되게 산출되는가의 문제일 것이다. 2005학년도 수능 이후, 1등급이 과다 양산되어 2등급이 산출되지 않는 현상이 일부 과목에서 있었다. 너무 쉽게 출제되거나 수험자 집단의 능력 분포가 매우 특이한 경우 나타날 수 있는 현상이다. 현재까지는 표준점수나 백분위가 제공되었기 때문에 등급 점수의 문제점을 일부나마 보완해줄 수 있었다. 하지만 등급만 제공되는 2008학년도 이후 수능에서는 출제 과정과 점수 체제 내에서만 이를 해결해야 하는 어려운 상황에 처하게 되었다.

이에 본 연구에서는 영역/과목별 안정적인 등급 산출을 위한 조건을 탐색하기 위하여 첫째, 검사 점수 유형에 따른 등급 산출 결과를 비교하였다. 즉, 등급 산출을 위해 가능한 검사 점수 분포로서 원점수, 선형 표준점수, 그리고 정규화 표준점수 분포를 검토하였다. 그리고 검사 점수 유형 이외에 문항 난이도가 등급 산출에 미치는 영향을 분석하였다. 즉, 기준 비율에 근사한 등급점수의 산출을 위한 검사 조건을 분석하기 위하여 영역/과목별 문항 난이도[2]의 변화와 등급점수의 관계를 분석하였다.

2) 고전검사이론에 근거한 문항 정답률로서 전체 피험자 중 정답한 피험자의 비율을 의미한다.

Ⅱ. 등급 산출 방식

1. 스테나인 방식

2008 대입 개선안에 따라 2008학년도 수능부터 현재의 표준점수와 백분위를 제공하지 않고 등급만을 제공하게 된다. 표준점수 혹은 백분위 점수의 1, 2점 차이에 민감해지는 치열한 점수 경쟁을 완화하기 위한 등급 점수의 활용에 대하여 교육 관계자와 관련 단체들이 대체적으로 동의함에도 불구하고 등급의 수와 산출방식은 상당한 논란의 여지가 있다. 즉, 과도하게 세분화된 등급은 현 수능 체제의 점수 경쟁을 완화할 수 없으며, 적은 수의 등급은 수능의 변별력을 약화시켜 대입전형자료로써의 활용도를 저해하게 된다.

2008학년도 수능부터 제공될 등급 점수는 지금의 그것과 동일한 스테나인 체제이다. 스테나인은 정규화 척도점수의 일종으로 제2차 세계대전 기간 중 미국 공군(United States Air Force)에서 개발한 점수 체제며, 주로 적성검사와 성취도 검사의 점수를 보고하는 데 이용되어 왔다(Anastasi & Urbina, 1997). 스테나인은 정해진 비율에 따라 등급 점수를 부여하는 방식으로 산출되며 원점수가 정규분포일 때 평균 5 표준편차 2의 특성을 갖는다. 또한 8개의 등급 분할점은 z 점수(평균 0, 표준편차 1의 표준점수) 상에서 약 0.5점(1/2 표준편차) 간격으로 동간성을 유지한다 ([그림 1] 참고).

스테나인은 다른 척도점수와 비교하여 많은 원점수를 하나의 척도점수로 전환한다는 면에서 정보의 양이 작아지는 단점을 갖는다. 그러나 9개의 등급 점수로 표현되는 스테나인은 단순성이라는 측면에서는 장점을 갖는다(Mehrens & Lehman, 1991). 또한 백분위 같은 세밀한 서열 점수와 비교하여 일정구간에 동일한 점수를 부여하는 등급 점수는 작은 차이에 대한 과장된 해석을 방지하는 역할을 한다(Crocker & Algina, 1986). 실제, 스테나인의 두 점수가 2 이상 차이가 나는 경우 두 점수 사이에 신뢰로운 차이가 있는 것으로 해석할 수 있다.

스테나인 1	2	3	4	5	6	7	8	9
4%	7%	12%	17%	20%	17%	12%	7%	4%
z점수	-1.75	-1.25	-0.75	-0.25	0.25	0.75	1.25	1.75
T점수	32	37	42	47	53	58	63	68

〔그림 1〕 스테나인의 등급 분할 점수

이상 논의된 스테나인의 이론적 특성은 원점수의 정규분포와 연속분포라는 두 가지 전제에 근거하는 것으로, 검사 점수에 스테나인의 등급 점수를 부여할 경우 다음과 같은 문제점을 가진다. 첫째, 수험자의 특성을 제한된 과제(일반적으로 문항)로 측정한 결과로 나타나는 검사 점수는 반드시 정규분포를 보이지는 않는다. 예를 들어 검사가 수험자의 평균 능력보다 쉬운 문항으로 구성된다면, 수험자의 능력이 정규분포일지라도 검사 점수는 부적편포를 보일 것이다. 이 경우 원점수를 스테나인으로 전환하게 되면, 상위 등급에 대한 분할점을 찾는 데 어려움이 나타날 것이며, 1/2 표준편차 동간성이라는 특성을 상실하게 된다. 둘째, 일반적으로 검사 점수는 연속분포가 아닌 이산분포를 갖는다. 정규분포는 연속분포를 가정한 이론적 분포로 양극단이 무한 값을 가지며, 필요에 따라 두 점수 사이를 무한히 분할하여 특정 백분위와 원점수를 짝지을 수 있다. 그러나 현실에 적용되는 검사 점수는 최고점과 최저점을 가지며, 많은 경우 정수 단위의 점수 체제를 갖는다. 이때 특정 정수 단위 분할점수에 많은 수험자가 밀집한 경우, 스테나인의 등급 당 비율을 정확히 유지하는 데 어려움을 보인다. 극단적인 예로, 한 검사에 대하여 만점자가 11%인 경우 1등급에 해당하는 수험자 비율은 11%가 되며 2등급에는 해당 수험자가 없는 경우가 나타날 수 있다. 이러한 이론적 정규분포와 현실적 검사 점수 분포의 차이는 이론과 현실의 간극이라는 측면에서 완전히 해결될 수는 없다. 그러나 검사의 난이도를 수험자 평균 능력 수준으로 조정하며, 문항 수와 배점을 조정하여 가능한 원점수의 가지 수를 많게 하는 방향으로의 노력 등이 이상의 문제점을 최소화하는 데 도움을 줄 것이다.

2. 다양한 등급 산출 방식

검사를 통하여 수험자의 특성을 나타내는 검사 점수는 다양한 형태를 가질 수 있다. 대표적으로 수험자 간의 서열 정보를 제공하는 규준참조점수(등수, 백분위, z점수, T점수, 스테나인 등)와 능력 수준에 대한 판단을 나타내는 준거참조점수(합격/불합격, 정답률) 등이 있다. 이러한 다양한 검사 점수 양식은 제공하는 정보가 서로 상이하며, 검사의 목적에 따라 장·단점을 가진다. 스테나인 또한 다양한 검사 점수의 하나로 수능이라는 검사의 목적과 활용이라는 측면에서 필요에 따라 적절히 변화/활용될 수 있을 것이다.

가. 수정된 스테나인 방식

스테나인의 대표적인 수정 방법은 'Rule of Four'이다. 이는 [그림 1]의 등급간 비율을 4% 단위로 단순화한 것이다. 즉, 각 등급의 비율이 4, 8, 12, 16, 20, 16, 12, 8, 4%로 설정된다. 이러한 점수 체계는 등급 비율 결정의 단순성에도 불구하고, 애초 스테나인의 동간격성, 평균 5, 표준편차 2 등의 특성을 상실하여 통계적 정확성이 떨어지게 된다. 또 다른 수정안은 스테나인의 통계적 정확성을 높이기 위하여 Kaiser(1958)가 제안한 것으로 5등급의 비율을 19%로 하고 1, 9등급을 4.5%로 조정한 것이다. 이러한 Kaiser의 수정된 스테나인은 정확히 평균 5 표준편차 2의 특성을 갖는다. 그러나 Kaiser의 방식은 정규분포의 수리식을 통한 이론적 논의에 제한된다.

나. 균일안

균일안은 스테나인 등급의 비율에 대한 논의로, 각 등급 비율을 균일하게(예, 11%) 하자는 주장이다. 스테나인의 8개 분할점은 서로 몇 표준편차 떨어져 있는가라는 의미에서 1/2 표준편차의 동간성이라는 점수설정의 규칙성과 해석상 일관성이라는 장점을 갖는다. 그러나 균일안은 등급비율에 대한 기계적 동등성에 제한된다. 또한 1, 2등급에 11%를 할당하는 경우 원래의 스테나인과 비교하여 상위권 수험자들에 대한 변별이 상대적으로 어려워진다.

3. 외국 사례

등급 점수로서 대부분의 표준화 검사는 스테나인을 이용한다(Hood & Johnson, 1997). 그러나 스테나인은 단독으로 사용되기 보다는 다른 검사 점수(표준점수, 백분위 등)와 함께 보고되는 것이 일반적이다. 등급 점수로서 일반적으로 이용되는 스테나인은 학생의 서열에 근거한 규준(norm)점수의 일종이라면 성취 수준에 따른 등급 설정은 준거(criterion)점수의 특징을 갖는다.

앞에서 살펴본 것과 같이 현행 수능에서는 영역/과목의 성격이나 수준이 다름에도 불구하고 일정 비율로 등급을 부여하는 규준참조 방식을 사용하고 있다. 남명호(2005)는 규준참조 방식에 의한 등급 부여의 문제점을 지적하고, 절대 기준에 비추어 등급을 부여하는 준거참조 방식에 의한 등급 부여 방법을 그 대안으로 제시하였다. 다음에서 현재 외국의 표준화 검사에서 등급 부여 방식을 간단히 살펴보았다.

가. 규준지향적 사례

Stanford 9 Achievement Test는 유치원에서 12학년을 대상으로 한 성취도 검사이다. 읽기, 수리, 언어 영역으로 구성되었으며, 백분위, 원점수, 학년점수 등과 함께 스테나인을 점수 체제로 활용한다. WIAT-Ⅱ(Wechsler Individual Achievement Test, Psychological Corporation, 2001)는 4세에서 85세까지를 대상으로 하는 성취도 검사이다. 읽기, 수리, 쓰기, 말하기 및 하위 9개 영역으로 구성되었으며, 백분위, 표준점수, T점수와 함께 스테나인이 점수 체제로 이용된다. Test of Cognitive Skills는 2학년에서 12학년을 대상으로 하는 지능 검사이다. 언어, 비언어, 단기기억에 관한 영역으로 구성되며, 백분위, IQ 점수와 함께 스테나인이 점수 체제로 이용된다.

나. 준거지향적 사례

영국 중등학교 졸업 자격시험(GCSE, General Certificate of Secondary Education)은 전기 중등교육까지 학생들의 교육성취도를 평가하고 그 도달 수준을 증명해주는 시험으로, 점수 체제는 A*, A, B, C, D, E, F, G, U 등의 9개 등급으로 이루어지며 각 등급은 학생들의 상대적 서열을 의미하는 것이 아니라, 지식과 이해, 혹은 기술과 능력의 성취 수준을 의미한다. 또한 영국 대학입학자격시험(GCE, General Certificate of Education)은 대학 진학시험의 역할을 하며 수준에 따라 A-level, AS-level, AEA 등 세 가지 시험으로 구성된다. 먼저 A-level(Advanced Level)은 고급 수준의 시험으로 과목별 시험으로 구성되며 10등급의 점수 체제를 갖는다. AS-level(Advanced Supplementary Level)은 A-level을 보충하는 시험으로 5등급 점수 체제를 갖는다. AEA(Advanced Extension Awards)는 A-level보다 우수한 학생을 대상으로 하는 시험으로 unclassified, merit, distinction 등 3단계의 점수 체제를 갖는다.

Ⅲ. 연구 방법

1. 연구 자료

본 연구에서 등급제를 논의하기 위하여 2005년에 실시된 두 번의 모의평가, 즉 2006학년도 수능 6월 및 9월 모의평가 자료를 이용하였다. 모의평가의 과목/문항 수, 시행시간, 시행방식은 본수능과 동일하며, 영역별 응시자 수는 <표 1>과 같다. 본 연구의 분석에는 언어, 수리 '나'형, 외국어(영어) 영역, 탐구 영역에서는 윤리, 제2외국어/한문 영역에서는 프랑스어I 과목이 포함되었다.

〈표 1〉 2006학년도 수능 6월 및 9월 모의평가 영역별 응시자 수(명)

구분	언어 영역	수리 영역	외국어 (영어) 영역	탐구 영역			제2외국어/ 한문 영역
				사회	과학	직업	
6월	582,579	555,614	581,694	305,366	191,728	81,729	55,981
				578,823			
9월	532,292	499,043	531,547	290,816	184,551	52,561	47,492
				527,928			

* 선택과목 간 분석결과로 제시한 윤리, 프랑스어I의 응시자 수는 6월 모의평가에서는 각각 155,087명과 4,099명, 9월 모의평가에서는 각각 138,864명과 3,725명임.

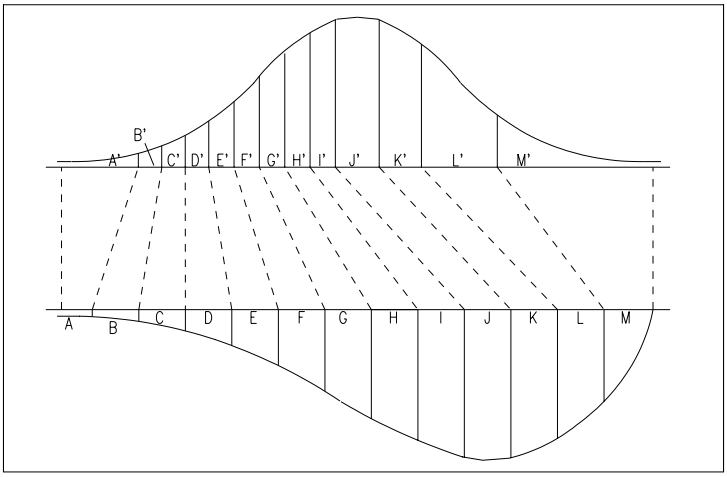
자료 분석에 이용된 2006학년도 6월, 9월 모의평가는 1학기와 2학기라는 시험실시 시기, 수시모집 응시자의 참여 여부 등으로 인하여 응시자 집단의 능력 수준에서 상당히 차이를 보인다고 할 수 있다. 응시집단의 능력 수준의 차이는 고전검사이론에 기초한 현 수능 점수체제의 중요한 고려사항으로 많은 선택과목 중 윤리와 프랑스어I은 이러한 편차를 경험적으로 보여주는 사례로 선택되었다.

2. 분석 절차

본 연구에서는 영역/과목별 안정적인 등급 산출을 위한 조건을 탐색하기 위하여 첫째, 검사 점수 유형에 따른 등급 산출 결과를 비교하였다. 즉, 등급 산출을 위해 가능한 검사 점수 분포로서 원점수(raw scores), 선형 표준점수(linearly transformed standard scores), 그리고 비선형 표준점수(nonlinear transformed standard scores) 분포를 검토하였다. 여기에서 원점수는 피험자가 맞힌 문항 점수(배점이 고려됨)의 합으로 정의된다. 따라서 언어, 수리, 외국어(영어)

영역의 원점수 범위는 0점～100점, 탐구 및 제2외국어/한문 영역은 0점～50점이 된다. 한편, 선형 표준점수는 원점수 분포를 일정한 평균과 표준편차를 이용하여 선형적으로 변환한 점수를 의미한다. 2005학년도 수능부터 사용된 각 영역/과목별 평균과 표준편차를 그대로 이용하여 원점수를 변환하였다. 즉, 언어, 수리, 외국어(영어) 영역의 경우는 평균 100, 표준편차 20 그리고 탐구 및 제2외국어/한문 영역의 경우는 평균 50, 표준편차 10인 표준점수를 사용하였다. 이러한 선형 표준점수는 기본적으로 원점수와는 서로 일대일의 대응관계를 갖게 된다. 하지만, 현행 수능 점수체제에서는 소수 첫째 자리에서 반올림하여 정수화된 표준점수를 보고하고 있기 때문에, 일부 서로 다른 원점수가 동일한 표준점수로 변환되는 현상이 나타나기도 한다. 본 연구에서 사용된 선형 표준점수란 현행 수능 점수체제에서 사용하고 있는 표준점수, 즉 정수화된 표준점수 분포를 의미한다.

비선형 표준점수는 정규화 표준점수(normalized standard scores)라고 불리는데, 그 산출 절차는 다음과 같다. 먼저 각 원점수에 해당하는 누가 백분율(cumulative percentile)을 계산하고 정규분포와 z점수와의 면적 비율과의 관계에 따라 각 점수의 누가 백분율에 해당하는 z점수를 구한다. 이 때 해당 z점수가 각 점수에 대응하는 정규화 표준점수(normalized z-score)가 된다(Allen & Yen, 1979; Guilford & Fruchter, 1981). 이러한 정규화 표준점수 분포의 변환 절차를 도식화하면 [그림 2]와 같다. 즉, [그림 2]의 하단에 위치한 편포되어 있는 한 과목의 분포(A～M)가 상단의 정규 분포(A'～M')에 근거하여 면적 변환이 되는 과정을 예시하고 있다. 이처럼 한 점수 분포는 면적 변환과정을 통하여 일부 점수 구간을 좁히거나 늘어나게 함으로써 원래 점수 분포의 모양을 정규 분포의 모양에 근사하게 가도록 변환시키게 된다. 본 연구에서 사용한 비선형 혹은 정규화 표준점수도 소수 첫째자리에서 반올림한 정수화된 표준점수 분포를 사용하였다.



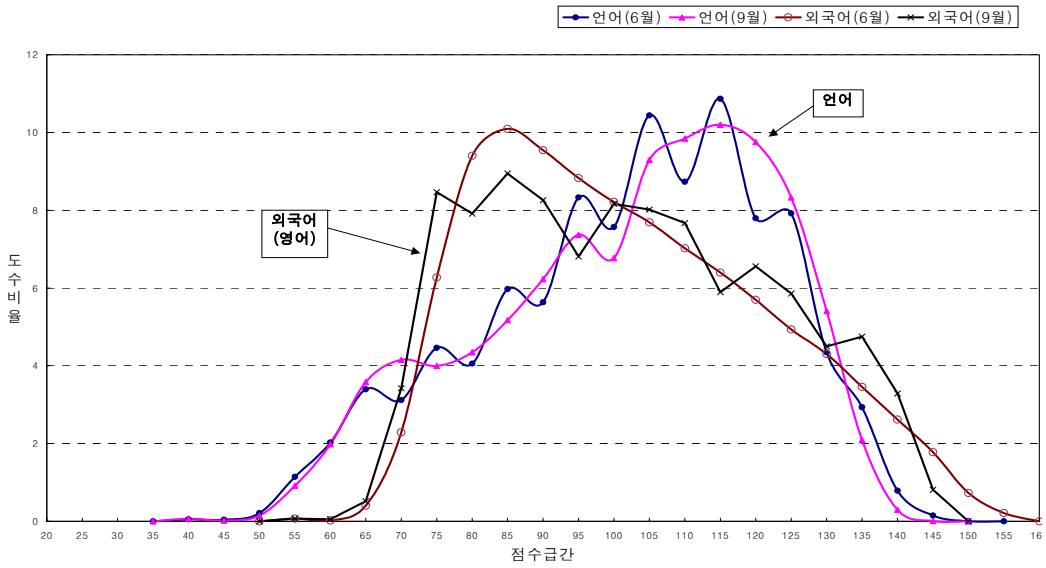
［그림 2］ 비선형(정규화) 표준점수 변환 예시

이러한 검사 점수 유형 이외에 영역/과목별 안정적인 등급 산출을 위한 조건을 탐색하기 위하여 문항 난이도가 등급 산출에 미치는 영향을 분석하였다. 즉, 목표 기준 비율에 근사한 등급점수의 산출을 위한 검사 조건을 분석하기 위하여 영역/과목별 문항 난이도와 등급 비율간의 관계를 분석하였다.

Ⅳ. 연구 결과

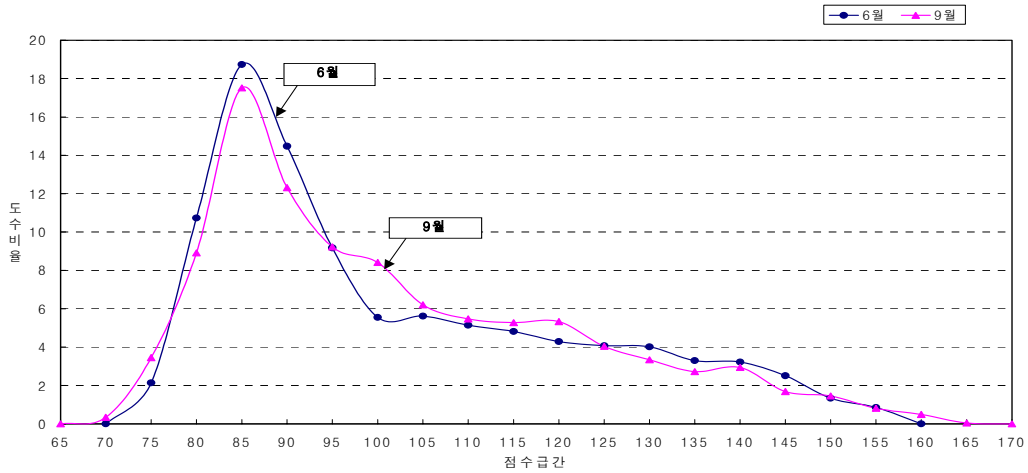
1. 검사 점수 분포 비교

검사 점수 유형 및 문항 난이도에 따른 등급 산출 결과를 비교하기에 앞서, 2006학년도 수능 6월 및 9월 모의평가의 영역/과목별 점수 분포의 특성을 검토하기 위하여 영역/과목별 표준점수(정수화된 선형 표준점수 분포) 분포를 비교하였다. 먼저 선택과목이 없고 대부분의 수험생들이 응시하고 있는 언어와 외국어(영어) 영역인 경우는 비교적 중앙 부분에 대다수의 수험생들이 위치하고, 양극단에 소수의 수험생들이 분포하는 형태를 나타내고 있다. 언어 영역의 경우는 다소 부적편포(negatively skewed)의 경향을 보이고 있고, 외국어(영어) 영역에서는 다소 정적편포(positively skewed) 형태를 띠고 있다.



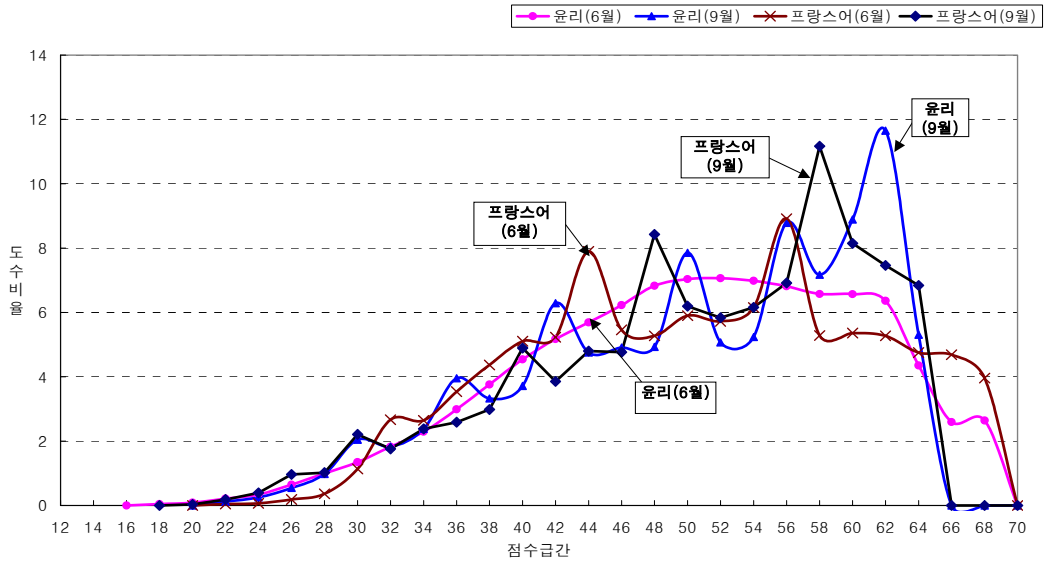
［그림 3］ 언어와 외국어(영어) 영역의 표준점수 분포

반면, 수리 '나' 형의 경우는 정적편포(positively skewed)를 나타내는 점수 분포로서, 상대적으로 하위권에 많은 학생들이 집중되어 있고, 상위권 학생들이 적은 분포형태를 보인다.



〔그림 4〕 수리 '나' 형의 표준점수 분포

다음으로는 선택과목이 있는 탐구 및 제2외국어/한문 영역의 점수 분포를 살펴보면, 선택과목별로 분포의 형태가 정규분포와는 다소 차이가 있을 뿐 아니라, 선택과목별로 그리고 시기별로(6월 또는 9월) 상이한 형태를 나타냈다. 예를 들어서, [그림 5]에 있는 사회탐구 영역의 윤리 과목의 경우 6월 모의평가에서는 검사 점수 분포가 대체적으로 정규분포에 근사하게 평균을 중심으로 좌우 대칭인 형태를 나타냈다. 하지만, 9월 모의평가의 경우는 검사가 쉽게 출제되어서 매우 부적인 편포를 나타내고 있다. 제2외국어/한문 영역의 프랑스어 I과목의 경우에도 6월 모의평가에서는 검사 분포가 양봉(bimodal) 형태를 취하기는 하지만, 비교적 전 능력대에 골고루 수험생들이 분포하고 있는 반면, 9월 모의평가에서는 부적인 편포 경향이 나타나고 있다.



〔그림 5〕 탐구 및 제2외국어/한문 영역의 표준점수 분포

2. 검사 점수 유형에 따른 등급 산출 결과 비교

2006학년도 수능 6월 및 9월 모의평가 자료를 이용하여 원점수, 선형 표준점수, 그리고 정규화 표준점수 분포상에서 스테나인 점수 즉, 등급을 산출하고, 그 결과를 비교하였다.

먼저, 언어, 수리 '나' 형, 및 외국어(영어) 영역의 경우는 검사 점수 유형별로 산출된 등급 비율 간에는 큰 차이를 나타내지 않았으며, 산출된 등급 비율은 스테나인의 기준 등급 비율과 비교적 근사한 패턴을 보였다. 다만, 수리 '나' 형의 경우는 원점수와 정규화 표준점수의 등급 비율은 유사한 패턴을 보인 반면 선형 표준점수에서는 다소 차이를 나타냈다.

〈표 2〉 언어와 외국어(영어) 영역의 검사 점수 유형에 따른 등급 비율 비교

등급	기준 비율	언어						외국어(영어)					
		6월 모의평가			9월 모의평가			6월 모의평가			9월 모의평가		
		원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수
1	4.00	4.87	4.87	4.87	4.48	4.48	4.48	4.43	4.43	4.43	4.06	4.06	4.06
2	7.00	6.75	6.75	6.75	8.25	8.25	8.25	6.74	6.74	6.74	7.39	7.39	7.39
3	12.00	12.87	12.87	12.87	12.36	12.36	12.36	12.24	12.24	12.24	11.62	11.62	11.62
4	17.00	16.62	16.62	16.62	16.33	16.33	16.33	17.81	17.81	17.81	17.26	17.26	17.26
5	20.00	19.21	19.21	19.21	18.89	18.89	18.89	20.28	20.28	20.28	20.46	20.46	20.46
6	17.00	17.28	17.28	17.28	16.76	16.76	16.76	17.03	17.03	17.03	16.41	16.41	16.41
7	12.00	12.01	12.01	12.01	12.12	12.12	12.12	10.74	10.74	10.74	11.87	11.87	11.87
8	7.00	6.55	6.55	6.55	7.29	7.29	7.29	7.35	7.35	7.35	7.37	7.37	7.37
9	4.00	3.83	3.83	3.83	3.52	3.52	3.52	3.39	3.39	3.39	3.57	3.57	3.57

〈표 3〉 수리 영역('나' 형)의 검사 점수 유형에 따른 등급 비율 비교

등급	기준비율	'나' 형					
		6월 모의평가			9월 모의평가		
		원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수
1	4.00	4.23	4.23	4.23	4.72	4.72	4.72
2	7.00	7.73	7.73	7.73	6.64	6.64	6.64
3	12.00	11.78	11.78	11.78	**11.64**	**12.57**	**11.64**
4	17.00	16.58	16.58	16.58	**17.27**	**16.35**	**17.27**
5	20.00	19.74	19.74	19.74	21.25	21.25	21.25
6	17.00	18.48	18.48	18.48	16.26	16.26	16.26
7	12.00	**11.82**	**13.84**	**11.82**	11.83	11.83	11.83
8	7.00	**5.92**	**3.90**	**5.92**	**7.27**	**8.23**	**7.27**
9	4.00	3.71	3.71	3.71	**3.11**	**2.15**	**3.11**

다음으로, 탐구 및 제2외국어/한문 영역의 경우는 2006학년도 수능 6월 및 9월 모의평가의 경우 [그림 5]에서 보듯이, 각 과목별 분포 모양이 서로 상이하여 산출된 등급별 비율과 기준 비율은 상당한 차이를 보였다. <표 4>에서는 사회탐구 영역의 윤리 과목, 제2외국어/한문 영역의 프랑스어 I 과목의 검사 점수 분포에 따른 등급 비율을 제시하였다. 대체적으로 원점

수와 정규화 표준점수 분포상에서 산출된 등급별 비율 분포는 유사한 패턴을 나타냈다. 반면, 서로 다른 원점수가 동일한 선형 표준점수로 변환된 경우, 일부 등급에서는 비율의 차이가 나타났다. 예를 들어, 윤리 과목(6월 모의평가)의 5등급 비율이 선형 표준점수 분포에서는 23.48%였으나, 원점수 및 정규화 표준점수 분포에서는 20.26%로 나타났다. 프랑스어I 과목(9월 모의평가)에서는 선형 표준점수 분포의 4등급이 19.95%였으나, 원점수 및 정규화 표준점수 분포에서는 16.40%이었다.

〈표 4〉 탐구 및 제2외국어/한문 영역의 검사 점수 유형에 따른 등급 비율 비교

등급	기준 비율	윤리						프랑스어I					
		6월 모의평가			9월 모의평가			6월 모의평가			9월 모의평가		
		원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수	원점수	선형 표준점수	정규화 표준점수
1	4.00	4.30	4.30	4.30	5.95	8.05	5.95	5.22	5.22	5.22	13.40	13.40	13.40
2	7.00	8.90	8.90	8.90	7.24	5.13	7.24	9.86	9.86	9.86	-	-	-
3	12.00	12.02	12.02	12.02	12.28	12.28	12.28	10.05	10.05	10.05	10.95	10.95	10.95
4	17.00	17.42	17.42	17.42	16.03	16.03	16.03	16.88	19.96	19.96	16.40	19.95	16.40
5	20.00	20.26	23.48	20.26	19.22	19.22	19.22	20.66	17.59	17.59	20.19	16.64	20.19
6	17.00	15.99	12.77	18.62	16.97	16.97	16.97	14.47	14.47	14.47	16.59	16.59	16.59
7	12.00	10.14	10.14	7.50	11.98	13.40	11.98	13.49	13.49	13.49	11.95	11.95	11.95
8	7.00	7.84	7.84	7.84	7.10	5.69	7.10	5.51	6.93	5.51	6.93	6.93	6.93
9	4.00	3.14	3.14	3.14	3.23	3.23	3.23	3.85	2.44	3.85	3.60	3.60	3.60

요약하면, 원점수, 선형 표준점수 및 정규화 표준점수 분포를 이용하여 등급 비율을 비교해 볼 때, 산출된 등급 비율은 영역/과목별, 그리고 검사 점수 유형별로 일부 차이가 있었다. 우선 영역/과목별 등급 비율의 안정성을 비교해 본 결과, 상대적으로 수험생의 수와 점수의 가지 수[3]가 많고, 검사 점수 분포가 대체로 정규 분포에 근사한 형태를 나타내는 언어, 수리 및 외국어(영어) 영역의 경우는 검사 점수 유형에 따른 등급별 비율에는 큰 차이를 나타내지 않았다. 또한 각 영역의 등급 비율 분포는 기준 비율에 근사하였고, 세 영역 간 등급 비율 분포도 비교적 유사한 패턴을 보였다. 반면, 탐구 및 제2외국어/한문 영역에서는 점수의 가지 수도 적을 뿐 아니라, 선택과목 별로 검사 점수 분포 모양이 서로 상이하여, 산출된 등급 비율 분포가 안정적이지 못함을 알 수 있었다.

3) 수능의 영역/과목별 원점수 가지 수는 문항수와 배점에 따라 결정된다. 예를 들어 60문항, 3가지 배점으로 구성된 언어영역의 원점수 가지 수는 101개(0∼100)이며, 20문항, 2가지 배점으로 구성된 사회탐구 과목의 원점수 가지 수는 49개(0∼50, 1과 49 제외)이다.

한편, 검사 점수 유형별 등급 비율 분포를 살펴보면, 대체로 원점수와 정규화 표준점수 분포는 유사한 결과를 나타낸 반면, 선형 표준점수 분포에서는 일부 등급 비율에 차이를 보였다. 이는 표준점수의 산출 과정에서 반올림 과정이 포함됨으로 인하여 서로 다른 원점수가 동일 표준점수로 변환되었기 때문이다. 이러한 현상은 정규화 표준점수 보다는 선형 표준점수 분포상에서 더 많이 나타남을 알 수 있었다. 따라서 등급 산출을 위한 가장 적절한 검사 점수 분포는 점수의 가지 수를 최대한 보장할 수 있는 분포로서, 본 연구 결과에 따르면 원점수 혹은 정규화 표준점수 분포였다. 하지만, 프랑스어I 과목(9월 모의평가)의 경우처럼, 만점자가 많이 발생하여 1등급의 비율이 기준 비율을 초과하고, 2등급이 산출되지 않는 등의 문제점은 세 가지 검사 점수 유형 모두에서 나타남을 알 수 있다. 이는 원점수 분포의 서열에 관한 정보는 해당 원점수를 선형 또는 비선형으로 변환시킨다 할지라도 그대로 유지되기 때문이다.

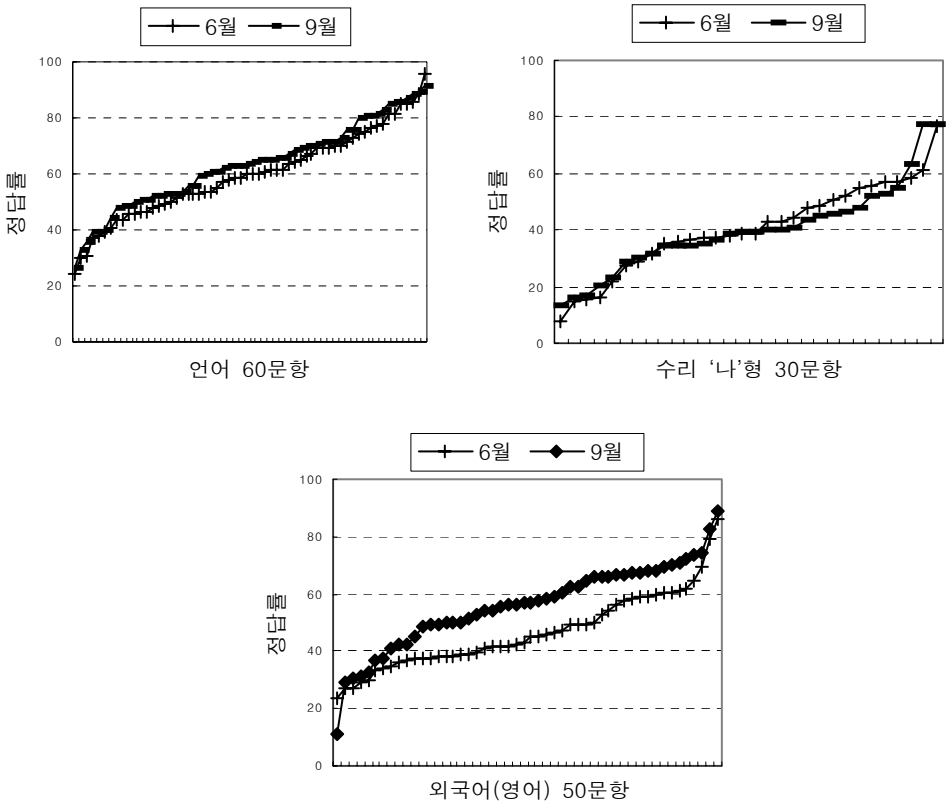
결론적으로, 많은 응시자를 변별해야 하고 상대적으로 점수 가지 수가 적은 시험 체제에서는 가능한 모든 점수 가지 수를 살릴 수 있는 분포를 이용하여 등급 점수를 산출하는 것이 가장 안정적임을 알 수 있다. 또한 등급은 영역/과목 내 서열 정보에 기반한다는 특성을 고려할 때, 원점수를 가능한 한 변환하지 않고, 원점수 분포상에서 등급 점수를 산출하는 것이 타당한 방안으로 보인다. 하지만, 동일 점수대에 많은 동점자가 발생한 경우에는 어떠한 검사 점수 분포에서도 이에 대한 보정은 불가능하며, 산출된 등급 비율 분포는 기준 비율과 차이를 나타낼 것이다. 따라서 검사 점수 분포가 가능하면 등급 산출의 전제 조건인 정규화 분포에 근사하게 될 수 있도록 출제과정에서의 보완이 선행되어야 할 것이다.

3. 문항 난이도에 따른 등급 산출 결과 비교

현 수능의 언어와 외국어(영어) 영역은 각각 60문항, 50문항의 상대적으로 많은 문항으로 구성되며 각 영역이 하나의 과목을 나타낸다. 그러나 수리 영역은 30문항이라는 다소 적은 문항 수를 갖는다. 대부분의 수험생이 응시한 언어, 수리 '나' 형, 외국어(영어) 영역에 대한 2006학년도 6월 모의평가와 9월 모의평가의 문항 정답률(난이도)의 분포는 [그림 6]에 제시되었다. 각 그림에서 문항의 순서는 실제 문항번호와 상관없는 문항 정답률 순이다.

두 번의 모의평가에서 안정적인 등급 점수(목표비율에 근사한 등급비율)가 산출된 언어, 외국어(영어) 영역의 문항 정답률은 20~90% 구간에 골고루 분포하였다. 이는 쉬운 문항, 중간 수준 문항, 어려운 문항 등 다양한 수준의 난이도를 갖는 문항으로 검사가 구성될 때 9개의 등급 점수가 안정적으로 산출되고 있음을 경험적으로 보여 주는 것이다. 그러나 수리 '나' 형의 경우 문항 정답률이 일관되게 10~80% 구간에 분포하여 상대적으로 어려운 문항으로 구성되었다.[4].

또한 지난 두 번의 모의평가의 언어, 수리, 외국어(영어) 영역에서 주목할 점은 시험 시기별 문항 난이도에 큰 편차가 없이 일관된 수준의 검사 난이도를 보여 준다는 것이다. 외국어(영어) 영역의 경우 상대적으로 큰 편차를 보이고 있으나 [그림 3]과 같이 점수 분포에는 큰 영향을 미치지 않는 수준이다. 즉, 시기별로 응시자의 능력 수준이 상대적으로 다름에도 불구하고 언어, 수리, 외국어 영역은 사전검사(pre-test)라는 검증절차가 없는 현재의 출제 체제의 어려움 속에서도 일관된 문항/검사 난이도 수준을 유지하는 안정성을 확보하고 있다.



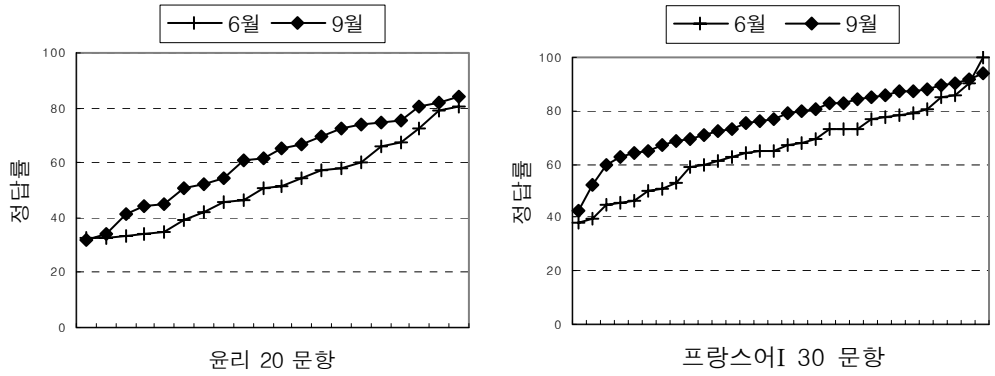
[그림 6] 언어, 수리 '나' 형, 외국어(영어) 영역 문항 정답률(** 점의 크기를 줄임)

수능의 사회/과학/직업 탐구 영역과 제2외국어/한문 영역은 각각 20문항, 30문항의 상대적으로 적은 문항으로 이루어진다는 점과 각 영역이 8~17 과목으로 구성되는 특징을 갖는다. 이러한 특성은 응시집단의 능력 수준이 시기별 과목별로 유동적임을 의미하며, 결국 동일

4) 수리 '나'형의 낮은 문항 정답률은 문항 자체의 난이도와 응시집단의 평균 능력 수준이라는 두 가지 요인이 복합적으로 작용한 것으로 이해될 수 있다.

과목 내의 시험 시기별 난이도 편차, 동일 영역 내 과목 간 난이도 편차, 불안정한 등급 점수의 산출이라는 문제점과 연결된다.

사회/과학/직업탐구 영역과 제2외국어/한문 영역의 2006학년도 6월 모의평가, 9월 모의평가에서 특정 시기에 불안정한 등급 점수를 나타낸 윤리와 프랑스어I의 문항 정답률(난이도)의 분포가 [그림 7]에 제시되었다. 윤리의 경우, 6월 모의평가에서는 1등급 비율이 4.30%로 목표기준 비율과 유사하게 산출되었으나, 9월 모의평가에서는 1등급 비율이 8.05%로 과도하게 산출되었다. 또한 프랑스어I의 경우, 6월 모의평가에서는 비교적 안정적으로 모든 등급 산출이 되었으나, 9월 모의평가에서는 2등급이 산출되지 않았다. [그림 7]에서 문항의 순서는 실제 문항번호와 상관없는 문항 정답률에 따른 순서이다.



〔그림 7〕 윤리, 프랑스어I의 문항 정답률

[그림 7]에 나타난 바와 같이 등급이 안정적으로 산출된 6월 모의 평가의 문항 정답률은 9월에 비하여 낮게 나타난다. 이러한 결과는 상대적으로 어려운 문항이 포함될 때 탐구영역과 제2외국어 영역의 9개의 등급 점수가 안정적으로 산출되고 있음을 경험적으로 보여주는 것으로 특히 어려운 문항(정답률 50% 미만의 문항)이 적은 경우 상위권에 대한 변별이 어려워져 1등급이 목표비율(4.0%)보다 과다 산출되는 결과로 이어진다. 또한 지난 두 번의 시험에서 주목해야 할 점은 시험 시기별 문항 난이도에 편차가 커 검사 난이도가 상대적으로 일관되지 못하다는 것이다. 이러한 과목 내, 과목 간 검사 난이도의 편차라는 문제는 상대적으로 적은 수의 문항으로 구성된다는 영역 특성과 관계된 것이다.

Ⅴ. 요약 및 결론

본 연구에서는 수능의 안정적인 등급 산출을 위한 조건을 탐색하기 위하여 검사 점수 유형과 문항 난이도가 등급 산출 결과에 미치는 영향을 분석하였다.

원점수, 선형 표준점수, 정규화 표준점수 분포를 이용하여 등급 비율을 산출했을 때, 점수의 가지 수가 많은 언어, 수리, 외국어(영어) 영역의 경우는 탐구 및 제2외국어/한문 영역에 비하여 검사 점수 유형에 따른 등급별 비율에 큰 차이가 없는 것으로 나타났다. 그리고 탐구 및 제2외국어/한문 영역의 선택과목과 같이 점수 가지 수가 적은 시험에서는 가능한 모든 점수 가지 수를 살린 분포, 즉 원점수 분포를 이용하여 등급 점수를 산출하는 것이 효과적인 것으로 나타났다.

문항 난이도와 등급 비율 간의 관계를 보면, 사전검사라는 검증절차가 없는 현재의 출제체제의 한계에도 불구하고, 언어, 수리, 외국어(영어) 영역은 매 시험마다 문항 정답률이 고르게 분포하여 안정성을 유지하고 있는 것으로 나타났다. 그리고 이 세 영역의 9개 등급 비율도 기준 비율에 유사하게 산출되고 있는 것으로 나타났다. 하지만, 문항 수(점수 가지 수)가 적은 탐구 및 제2외국어/한문 영역의 선택과목에 있어서는 문항 난이도 분포에 따라 등급 산출의 안정성이 영향을 받는 것으로 나타났다. 즉, 선택과목의 등급이 안정되게 산출된 경우에는 문항 정답률이 약 30~90% 구간에 골고루 분포하였던 반면, 2등급이 없어지는 경우에는 많은 수의 문항 정답률이 60% 이상이었다. 특히 어려운 문항이 없는 경우 상위권에 대한 변별이 어려워져 1등급이 과다 산출되고 2등급이 사라지는 현상이 나타났다. 이러한 결과는 쉬운 문항, 중간 수준의 문항, 어려운 문항 등 다양한 수준의 난이도를 갖는 문항으로 시험이 구성될 때, 9개 등급 비율이 안정적으로 산출될 수 있음을 시사한다.

이상의 연구 결과에서 볼 때, 일정 비율에 따라 9개 등급을 부여하는 현행 등급 체제를 그대로 적용하는 경우, 점수의 가지 수가 많은 원점수[5]를 사용하여 등급을 산출하는 것과 문항의 난이도가 고르게 나오도록 출제 과정을 강화하는 것이 안정적인 등급 산출의 주요 요건이라 할 수 있다. 그러나 문항 수가 20개밖에 되지 않고 수험자 집단이 안정적이지 못한 선택과목의 경우에는 점수의 가지 수가 매우 적을 뿐만 아니라 문항의 난이도가 고르게 분포되도록 출제하는 것이 매우 어렵다는 시험 자체의 한계를 지니고 있어 매번 시험에서 9개 등급의 안정적 산출을 보장하기는 현실적으로 어려운 일이다. 따라서 현행 등급 체제를 그대로 유지하는 수능 체제하에서는, 단기적으로 앞서 말한 출제 과정상의 난이도 조정 노력을 통하여 안정적인 등급 산출을 유도해야 하고, 중·장기적으로는 선택과목의 통합·조정을 통해 시험과목의 수를 축소하고 문항 수를 확대 하는 방향으로 시험 체제를 개선해야 할 것이다. 또한 문제은행식 출제 방식은 사전 문항 정보를 수집하여 검사를 구성할 수 있다는 점에서 안정적 수능 출제와 점수 산출을 위해 실천적인 노력이 집중되어야 할 영역이다.

---

5) 선택과목 간 점수를 맞비교하는 경우에는 난이도 및 수험자 집단 특성을 고려한 표준점수 체제가 원점수 체제보다는 합리적이라 할 수 있다. 하지만, 2008학년도 수능부터는 단순히 특정 과목에서의 상대적 서열에 따른 등급만 제공하기 때문에, 굳이 표준점수로 전환할 필요가 없게 된다.

참 고 문 헌

교육인적자원부(2004). **학교교육 정상화를 위한 2008학년도 이후 대학입학제도 개선안**. 교육인적자원부.

남명호(2005). 2008학년도 이후 대학수학능력시험 개선 방안에 대한 비판적 검토. **교육평가연구, 18**(2), 17-33.

남명호 · 김신영 · 남현우 · 권순달 · 박정 · 조지민(2002). **2005학년도 대학수학능력시험 점수체제 연구**. 한국교육과정평가원 연구보고 CAT 2002-32.

한국교육과정평가원(2005). **대학수학능력시험 10년사**. 한국교육과정평가원.

Allen, M. J. & Yen, W. M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadsworth.

Anastasi, A. & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Holt, Rinehart and Winston.

Guilford, J. P. & Fruchter, B. (1981). *Fundamental Statistics in Psychology and Education*. Tokyo, Japan: McGraw-Hill.

Hood, A. B. & Johnson, R. W. (1997). *Assessment in Counseling: A Guide to the Use of Psychological Assessment Procedures* (2nd ed.). Alexandria, VA: American Counseling Association.

Kaiser, H. F. (1958). A modified stanine scale. *Journal of Experimental Education, 26*.

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology* (4th ed.). Austin, Texas: Harcourt Brace College Publishers.

Psychological Corporation (2001). *Wechsler Individual Achievement Test* (2nd ed.). (WIAT II).

• *논문접수 : 2006년 4월 12일 / 수정본 접수 : 2006년 5월 15일 / 게재 승인 : 2006년 5월 24일*

ABSTRACT

Exploratory Analysis for Stable Stanines of College Scholastic Ability Test

Kil-Seok Yang(Research Fellow, Korea Institute of Curriculum & Evaluation)
Kyung-Seok Min(Research Fellow, Korea Institute of Curriculum & Evaluation)
Won-Sook Sohn(Research Fellow, Korea Institute of Curriculum & Evaluation)
Myung-Ae Lee(Research Fellow, Korea Institute of Curriculum & Evaluation)

While current CSAT(College Scholastic Ability Test) score reporting includes percentiles and standardized scores as well as stanines, Minister of Education and Human Resource announced that only stanine score will be reported from 2008 school year CSAT. The purpose of this study is to evaluate several conditions for stable stanines. To complete this goal, we compared stanine proportions according to raw scores, linear standardized scores and normalized standard scores. Also we analyzed effects of item difficulties on stanine scores. We found that there were not much differences of stanine proportions in Korean Language, Mathematics and Foreign Language (English) whether we used raw scores, linear standardized scores or normalized standard scores. But raw scores made stanines more stable in Social Studies/Science/Vocational Education and Foreign Languages/Chinese Characters and Classics since raw scores had more score points compared with linear standardized scores and normalized standard scores. Also we found that stanines approached target proportions when a test included various item difficulty levels(i.e., difficult, moderately difficult and easy). In sum, CSAT stanine scales would be stable when a test consists of various levels of item difficulty and raw scores are used for the final scaling. But this suggestion may not be feasible when a test includes a relatively few items (i.e., 20 items for Social Studies). So it needs to be considered to increase the number of items by combining similar subjects in future.

Key Words : stanines, test score types, item difficulty