교육과정평가연구
The Journal of Curriculum and Evaluation
2010, Vol. 13, No. 2, pp. 175～195

국가수준 학업성취도 평가에 기초한 학교평가 통계 모형의 고찰

최 길 찬(CRESST/UCLA 연구원)*
김 성 열(한국교육과정평가원 원장)**

≪ 요 약 ≫

이 글은 최근 한국에서 뜨거운 감자로 부각되어 있는 교육이슈 중 하나인 국가수준 학업성취도 평가 자료를 어떻게 분석하고, 그 결과를 어떻게 보고하고, 또 활용할 것인가 하는 문제의식에서 출발하였다. 이러한 문제의식에서 이 글은 국가수준 학업성취도 평가에 기초한 학교평가의 방향, 평가 설계, 평가 지표에 대한 이해, 평가를 위한 검사척도의 문제, 통계적 모형 등을 종합적으로 고찰하였다. 성장평가 모형을 같은 학생을 반복적으로 측정한 자료를 이용하는 종단연구설계 모형과 같은 학년의 학생을 반복적으로 측정하는 코호트 성장모형으로 구분하고, 각각의 통계모형을 제시하였다. 현행 국가수준 학업성취도 평가는 기본적으로 같은 학년의 학생을 반복적으로 측정하는 코호트 설계에 따른 평가방법이다. 하지만, 수년 후에는 한 학생의 학업성취도의 변화를 종단적으로 추정할 수 있으며, 이러한 학생의 학업성취도의 변화에 학교급별로 - 초등학교, 중학교, 고등학교- 어떤 영향을 주었는지를 통계적으로 모형화 할 수도 있다.

주제어 : 국가수준 학업성취도 평가, 학교 책무성 평가, 현시점 평가, 성장 평가, 코호트 성장모형, 중다코호트 개인성장모형

Ⅰ. 시작하는 말

최근 각국의 교육개혁의 동향은 교육경쟁력을 높이고, 교육의 공적 책무성을 강조하는 방향으로 나아가고 있음을 보여주고 있다. 이러한 교육개혁의 흐름에는 교육은 흔히 국가의 백년지대계라는 말로 표현되듯 국가의 변화 에이전트를 육성하는 가장 큰 공적 사업 중 하

* 제1저자
** 교신저자, kim@kice.re.kr

나라는 인식이 전제되어 있다. 예컨대, 미국의 경우 2002년 제정된 “교육낙오자 방지법(No Child Left Behind: 이하 NCLB)” 법령의 공표와 함께 보다 강화된 학교 책무성 평가 체제를 지향하고 있다. 이러한 경향성은 오바마 정부가 출범하고 나서도 식을 줄 모르고 진행되고 있으며, 어떤 측면에선 더욱 강조되는 추세이다. 한국의 경우도 크게 다르지 않다. 이명박정부가 추진하고 있는 ‘기초학력은 국가가 책임진다’는 정책이 그 단적인 예이다. 특히, 교육성취도 관련 정보공개가 확대되면서, 최근 대학수학능력 시험 자료 공개에 따른 학교 간 격차 문제와 2년째를 맞고 있는 국가수준 학업성취도 평가자료에 대한 분석 및 활용방안이 가장 뜨거운 감자로 부각되고 있다. 그럼에도 불구하고, 그동안의 논의는 국가수준 학업성취도 평가 자체의 기술적 개선에 초점이 맞추어지고(김성열 · 남명호 · 정은영 · 김성숙, 2009), 국가수준 학업성취도 평가자료에 대한 분석 및 활용방안에 대한 논의는 드물었다.

이러한 맥락에서 이 글은 국가수준 학업성취도 평가에 기초한 학교평가의 방향, 평가 설계, 평가 지표에 대한 이해, 평가를 위한 검사척도의 문제, 통계적 모형 등을 종합적으로 고찰한다. 미리 밝혀 둘 것은 이 글에서 제시된 국가수준 학업성취도에 기초한 학교평가는 학생의 성취도, 즉 교육의 결과적 측면에 기반한 것이라는 점이다. 이 글은 학업성취도 평가 결과의 해석과 그것을 학교평가에 활용하는 것과 관련하여 시사점을 제공해 줄 것이다.

Ⅱ. 학교평가(monitoring school performance)의 시각

그럼, 먼저 학교평가에 대한 여러 가지 시각을 목적, 내용, 평가대상, 평가 유형 및 시점, 검사척도의 관점에서 검토해 보자. 모든 계획된 행위에는 목적이 있듯이, 학교평가의 목적 역시 크게 형성적(formative)목적과 학교의 교육책무성을 평가하는 총합적(summative) 목적으로 나눌 수 있다. 학교평가가 이 두 가지 목적 중 전적으로 하나만을 수행하는 것은 아니며, 두 개의 목적 중 어느 하나가 더 강조되는 형태라고 보는 것이 보다 타당할 것이다.

형성적 목적의 학교평가에서는 학생의 성취도 검사를 통해서, 학생이 어떤 내용을 숙달하였으며, 어떤 내용은 모르고 있는지, 모르면 왜 모르는지에 대한 정보 등 학생의 학습정보 수집이 주된 활동이다. 교사는 형성평가 결과를 활용하여 자신의 수업 계획을 세우고, 학생의 학습을 증진시키기 위한 수업 전략을 수립하게 된다. 학교의 입장에서 보면, 이러한 정보에 기반해서 학교가 가진 재정 및 자원을 보다 효율적으로 학생의 학업을 증진시키는 방향으로 사용할 수 있다. 학교평가가 이렇게 형성적 목적으로 이용되기 위해서는, 보통의 경우, 학생의 학업 달성 정도의 진단 및 이러한 정보 활용이 일련의 수업상황과 연계되어 아주 잦은 빈도로 이루어져야 한다. 따라서 평가문항의 수, 평가문항의 형태(사지 선다형, 단답형,

서술형), 평가결과의 신속한 처리, 평가결과를 통한 학생의 진단적 정보 획득, 평가결과에 따른 교사의 수업의 변경 및 적용의 계획이 일련의 연속적이고 체계적인 계획하에서 이루어져야 한다.

이에 반해, 교육책무성 평가를 위한 총합적 목적의 평가는 평가결과를 기초로 교사 혹은 학교의 수행수준을 판단하며, 그 수행수준에 따라 책임을 묻는 것이다. 이러한 목적이 교사나 학교에 공정하게 적용되기 위해서는 평가문항이 학교에서 가르치는 교육과정과 잘 부합해야 하며, 특히 총합적 평가의 결과는 학교 내에서 이루어진 학생의 학습에 기초해야 한다. 만일 학생의 학습활동이 사교육 등으로 인해 학교 밖에서 많이 이루어지고, 이러한 학습의 결과가 총합적 평가에 나타난다면, 이를 기초로 한 교사 혹은 학교의 책무성 평가는 그 타당성이 위협 받을 수 있다.

다음으로, 평가에 대한 두번째 측면을 평가 내용면에서 살펴보자. 학교도 다른 사회조직과 마찬가지로 하나의 유기체적인 특성을 가진 조직이라고 할 수 있다. 학교를 이 같은 조직적 관점(systems approach)에서 평가할 때 가장 흔히 생각해 볼 수 있는 것은 투입(input), 과정(process), 결과(output)에 대한 평가이다. 투입 부문에서는 학교의 예산, 시설, 설비 등을 포함하는 하드웨어적인 투입뿐 아니라, 학생의 입학 당시의 수준을 알 수 있는 학업 성취도 성적, 교사의 질적 수준(졸업 학교, 최종 학위, 교직 경력 등)을 포함한 소프트웨어적인 측면도 포함한다.

과정적 측면은 조직의 의사소통, 리더십, 조직원들 간의 관계, 의사결정 과정 및 조직원의 민주적인 참여, 학부모의 참여, 교사의 교수 과정의 질 등을 포함한 조직 전체의 건강도 체크가 그 주된 평가 대상이다. 이러한 조직적 측면은 학교의 투입적 측면과 결과적 측면을 연계하는 역할을 수행한다고 가정되며, 투입적 측면에서의 단점을 극복하고, 장점을 극대화하여, 높은 수준의 결과를 산출해 내는데 핵심적인 기능을 가진다고 할 수 있다.

결과적 측면은 학생의 학업성취도 측면과 비 학업성취도 측면으로 구분해 볼 수 있다. 학업성취도는 학교교육의 가장 중요한 산출물 중 하나로 간주되며, 학교의 책무성 평가에 가장 빈번하게 이용된다. 이 이외에도 학생의 졸업여부, 출석일 등도 학업성취도에 연관된 결과 변인으로 간주된다. 비 학업성취도 측면에서는 자아개념, 적성, 흥미, 사회성 등을 포함한 학생의 정의적 및 사회적 측면에서의 발달이 강조된다.

셋째, 평가의 또 다른 측면으로 평가대상을 생각해 볼 수 있다. 평가의 대상은 학생, 교사, 학교로 구분해 볼 수 있는데, 많은 경우에 학생을 대상으로 한 결과를 교사 수준으로 통합해서 교사 평가가 이루어지고, 학교 내 학생 전체를 통합해서 학교 평가가 이루어진다. 다시 말해, 교사 자체를 대상으로 자료를 수집하거나 학교를 하나의 단위로 자료를 수집하여 평가를 하는 경우를 제외하고는 학생을 대상으로 한 평가결과를 교사나 학교 단위로 총합해서 평가가 이루어진다.

넷째, 평가시점에 따른 유형 또한 중요한 요소 중 하나이다. 평가시점은 크게 현재 시점에서 현재 상태를 평가하는 현시점 평가와 종단적으로 여러 시점에서 평가하는 성장평가로 나누어 볼 수 있다. 성장평가는 다시 같은 학생을 반복적으로, 즉 여러 학년에 걸쳐 종단적으로 평가하는 방법과 같은 학년을 반복적으로 여러 해에 걸쳐 평가하는 방식으로 다시 나누어 볼 수 있다. 참고적으로 이 두 가지 방법을 함께 다 고려하는 평가설계도 가능하며, 현재 미국에서 시행하고 있는 NCLB 정책이 대표적인 예라고 할 수 있다.

마지막으로 검사척도는 평가결과의 활용과 의미해석상 중요한 요소이다. 가장 흔히 쓰이는 척도점수는 백분위 점수(percentile score), 표준점수(standardized score), 정상분포변환점수(Normal Curve Equivalents: 이하 NCE), 척도점수(scale score)이다. 각 척도점수는 서로 다른 특징을 가지고 있다. 백분위 점수는 가장 이해하기 쉬운 점수여서 학부모, 교사 및 교육행정가, 정책입안자 등이 선호하는 척도이며, 표준점수는 동간척도로 이루어져 있기 때문에 수리적인 연산이 가능할 뿐 아니라 척도 내에서의 상대적 위치 또한 쉽게 파악할 수 있는 장점이 있다. 척도점수는 점수의 활용 목적에 맞게 변환된 점수이며, 동간성을 가정할 뿐 아니라 척도상의 증감이 절대적 의미를 가진다고 가정된다. 평가척도에 대한 보다 자세한 논의는 이 글의 다른 절에 제시하기로 한다.

이 글은 학교평가의 다양한 측면 중에서 국가적 수준의 학업성취도 평가에 기초한 학교평가 모형에 대해 살펴본다. 학업성취도 평가는 국가적 수준의 학력의 변화를 측정하고, 학교의 책무성 평가를 위해 이용될 수 있는 총합적 목적의 평가라고 할 수 있다. 또한 평가내용적 측면에서는 결과에 초점을 맞춘 평가이며, 학생의 성취도를 측정하여, 이를 학교수준으로 총합하는 평가 방식이다. 다시 말해, 현행 국가수준 학업성취도 평가는 학생 개개인의 변화 혹은 성장을 측정하기보다는 학교의 책무성 및 학교효과를 종단적으로 측정하기 위해 고안된 평가체제이다.

이 점에서 국가수준 학업성취도 평가결과는 “이 학교가 저 학교보다 더 좋은 학교인가?” 혹은 “이 학교가 저 학교보다 학교효과가 큰가?” 라는 질문에 대한 대답을 제공할 수 있어야 한다. 이 질문은 언뜻 보기에 간단히 답변될 수 있는 것처럼 보인다. 하지만, 이러한 질문에 적절하게 답하기 위해선 다음의 질문에 먼저 답할 수 있어야 한다.

- 검사의 타당도 및 신뢰도가 보장되는가?(이는 검사를 통해 측정하려고 하는 “학력”이 정확하고 신뢰롭게 측정되어야 할 뿐 아니라 이를 근거로 해서 결론내리는 추론(inference)이 타당해야 함을 의미한다.)
- 학생을 단위로 한 측정의 결과를 학교단위로 총합할 수 있는가?(이는 검사가 학생의 “학력”을 측정한다고 할 때, 이를 학교단위로 총합하는 것이 학교의 수행 혹은 질을 측정하는데 신뢰롭고 타당한가라는 질문이다.)

- 학생의 변화는 어떻게 측정하는가? 학교수행수준 및 학교의 질이 변화한다고 가정할 때 어떤 방식으로 측정할 수 있는가?(검사 결과의 연도별 학년별 변화가 단지 검사가 다르고, 검사 척도가 달라서 나타나는 현상일 수 있다.)
- 수행표준(performance standard)이 절대적인 기준에 따라 정해진 것인가 아니면 상대적인 기준에 의해 정해진 것인가?(학생이나 학교의 수행수준을 비교 판단하기 위해서 만들어진 수행표준이 학생 간 혹은 학교 간의 상대적인 순위만을 나타낼 수도 있다. 이와 다르게 각 학년별로 절대적 기준에 의해 정해진 표준은 학생 혹은 학교의 수행수준이 이러한 절대적 표준에 도달했는지 여부를 판단하는 근거를 제공한다.)

다음 절에서 제시되는 내용들은 위에 제시된 질문들에 대한 답을 제공한다.

Ⅲ. 학교평가를 위한 검사척도

학생의 학업성취도 자료는 학교평가를 위한 가장 중요하고 기초적인 자료이다. 이를 기초로 한 학교평가의 성패는 데이터로 제공되는 학업성취도 자료의 질에 의해 결정된다고 해도 과언은 아니다. 검사문항의 적절성, 신뢰도, 검사결과 추론의 타당성 등이 중요한 요소임은 앞서 이미 지적하였다.

검사결과 활용 측면에 있어서 중요한 요소는 검사가 어떤 척도로 되어 있는가 하는 것이다. 검사결과를 기초로 학교평가를 한다고 할 때, 가장 흔히 사용되는 척도는 원점수, 정답률, 순위점수, 표준점수, 동등화된 척도점수(vertically equated scale score) 등이다. 이러한 척도의 질은 사용 목적에 따라 달라질 수 있는데, 가장 중요한 것은 학교평가를 위한 척도로 어떤 것이 어떤 이유에서 적절하고, 어떤 이유에서 적절하지 못한가 하는 것을 정확히 구별하는 것이다. 더 나아가 어떤 척도가 앞 절에서 제시된 질문에 답하는 데 과연 적절한 것인가를 고려해야 한다.

1. 순위점수(percentiles)

순위점수는 아마도 가장 이해하기 쉽고, 대중에게 잘 알려진 척도라고 할 수 있다. 미국의 SAT(Scholastic Achievement Test)나 GRE(Graduate Record Examination)가 이 순위점수를 이용한다. 순위점수의 가장 큰 특징은 학생의 수행수준을 전체집단 중의 순위로 나타낸다는 데 있다. 즉, 순위점수가 85라면, 그 학생은 전체규준집단의 85%의 학생들보다 수행수준이 높다는 의미이다. 순위점수의 가장 높은 점수인 99는 이 점수를 받은 학생의 수행수준이 전체규

준집단 내의 99%의 학생들보다 수행수준이 높다는 의미이다. 따라서 순위점수는 학생이 전체규준집단 내에서 몇 등을 했는지에 대한 정보를 제공한다.

그러나, 이처럼 해석상의 용이성과 어떤 의미에선 점수를 받은 당사자 혹은 학부모가 가장 원하는 정보를 제공한다고 할 수 있는 순위점수는 학교평가를 위한 척도로 적절하게 활용되기에 어려운 단점이 있다. 순위점수는 동간척도에 기반한 척도가 아니다. 즉, 한 범위에 있는 두 점수 간의 척도상의 거리는 다른 범위에 있는 두 점수 간의 척도상의 거리와 같지 않다. 예를 들어, 순위점수 50에서 51간의 차이는 91에서 90간의 차이와 같지 않다. 즉, 이들 두 점수 간의 차이는 모두 1이지만, 척도상의 거리에 있어서 90에서 91간의 거리가 50에서 51간의 거리보다 더 크다. 따라서, 같은 1점의 차이라도 90에서 91로의 변화가 50에서 51로의 변화보다 더 의미있다고 할 수 있다. 결론적으로 동간성에 기초하지 않은 순위점수를 학교수준으로 총합한 순위점수 학교평균은 부적절하다고 할 수 있다(Russell, 2000). 하지만, 어떤 순위점수 이상에 있는 학생의 퍼센트가 어느 정도인가 하는 질문에 대한 자료는 여전히 흥미로운 학교평가의 정보라고 할 수 있다.

2. 정상분포변환점수(Normal Curve Equivalents: NCE)

NCE는 순위점수를 동간척도인 정상분포점수로 변환시킨 척도이다. 따라서 비동간성 척도인 순위점수보다 한 단계 개선된 척도체계라고 할 수 있다. 동간성이 가정된 척도이기 때문에 학교 평균을 구한다든지 하는 산술적인 계산이 가능한 척도체계이다.

이 척도는 평균이 50, 표준편차가 21 이며, 정상분포를 이루기 때문에 개별 점수가 분포내에서 어디에 위치하는지를 쉽게 파악할 수 있다는 장점이 있다. 하지만, NCE에 따른 학생 혹은 학교 간의 비교는 규준집단 내에서의 상대적인 의미만을 내포한다. 이러한 특징은 학생 혹은 학교를 종단적으로 평가할 때 중요한 의미를 가진다. 어떤 학생의 초등학교 3학년부터 6학년까지의 NCE로 나타낸 국어 학업성취도 점수가 동일하다고 가정해 보자. 이 경우, 이 학생의 국어 학업성취도 점수는 초등학교 4년 동안 전혀 변화하지 않았지만, 이것이 그 학생의 국어 “학력”이 증가하지 않았다고는 할 수 없다. 왜냐하면, NCE 점수는 규준집단 내에서의 상대적인 의미에 대한 정보만을 제공하기 때문에, 이 학생은 4년동안 상대적 위치에 있어서 전혀 변화가 없었지만, 그것이 실제 국어 학력이 증가하지 않았다고 말할 수 있는 근거가 되지는 않기 때문이다.

3. 동등화된 수직 척도 점수(vertically equated scale score)

척도점수는 원점수를 어떤 척도로 변환시켜 놓은 점수를 의미한다. 일반적으로 척도점수

는 서로 다른 검사에서 나온 점수들을 공통척도(common scale)에 배치시킨 점수이며, 이러한 특징으로 인해 학생 간 혹은 학교 간의 수리적 비교가 가능케 한 점수이다. 예를 들어, 올해 척도점수 650점을 받은 3학년 학생은 작년에 650을 받은 3학년 학생과 동등한 학력을 가진 것으로 판단할 수 있다. 이처럼 동일 학년 간 비교가 가능한 척도점수를 다른 학년들 간의 비교 또한 가능하도록 한 점수가 동등화된 수직척도점수이다. 이 점수는 3학년 척도에서의 5점의 증가가 5학년 척도에서의 5점의 증가와 동일한 의미를 가진다.

이러한 특징 때문에 동등화된 수직척도점수는 다년간의 학생의 학업성취도 변화를 측정하는 데 가장 적합한 척도이다(Hambleton & Swaminathan, 1987). 동등화된 수직 척도상에서의 연도별 점수의 변화는 상대적인 의미가 아닌 절대적 의미에서의 학력의 변화라고 할 수 있다. 동등화된 수직 척도점수를 얻기 위한 검사 동등화는 주로 공통문항을 여러 시점의 검사에 포함시켜 이루어지며, 이러한 절차는 검사 개발단계에서 어떤 문항을 포함시킬 것인가에 대한 구체적이고 체계적인 계획이 필요하며, 검사 실시 이후에도 척도를 동등화시키기 위한 기술적인 노력이 많이 요구된다. 동등화된 수직 척도점수가 가진 여러 장점에도 불구하고, 이러한 척도제작의 어려움으로 인해 미국의 52개 주 중에서 아주 소수의 주만이 3학년부터 8학년까지의 주 학력검사가 동등화된 척도점수로 되어 있다.

Ⅳ. 현시점(status)평가 대 성장(growth)평가

학생과 학교평가에 있어서 가장 중요한 요소 중 하나가 평가를 횡단적인 관점에서 실시하느냐, 아니면 종단적인 관점에서 실시하느냐 하는 것이다. 이 두 가지 관점은 평가를 하는 자료의 시점상의 횟수와 관련이 있을 뿐 아니라 측정의 단위인 개별 학생을 반복적으로 측정한 결과를 이용하느냐 여부와도 관련이 있다.

먼저, 이해가 쉬운 횡단적 관점에서의 학교평가에 대해 살펴보자. 과거 한국에서 실시된 국가수준 학업성취도 평가는 전집을 대상으로 하지 않고, 전집에서 5% 만을 표집하여 이루어졌다. 뿐만 아니라, 매년 표집되는 학교가 바뀌었다. 이 때, 개별 학교의 입장에서 보면, 학생의 학업성취도에 기초한 학교평가는 일회적인 것이었다고 할 수 있다. 이처럼 횡단적 평가는 평가가 일회적이란 것이 가장 큰 특징이다.

횡단적 평가는 일회적인 속성을 가지고 있는 반면에 몇 가지 장점 역시 가지고 있다. 가장 큰 장점은 자료수집의 용이성이다. 자료를 여러 시점에 걸쳐 종단적으로 수집할 때 가장 큰 어려움은 여러 가지 이유로 자료의 결측치(missing data)가 생길 가능성이 크다는 점이다. 상대적으로 횡단적 자료는 이러한 문제점이 적다고 할 수 있다. 두 번째 장점은 시간에 따

른 변화를 측정하기 위한 검사척도 개발에 대한 부담이 없다는 점이다. 많은 경우, 횡단적 자료에 기초한 학교효과 연구나 학교평가 연구가 학업성취도 자료와 함께, 방대하고 다양한 학생의 정의적, 인지적, 사회적 구인들 같은 개인 배경자료들을 측정하고, 이들 간의 상관관계를 포괄적으로 탐색할 수 있었던 것도 횡단적 자료수집 및 검사척도 개발의 상대적 용이성에 부분적으로 기인한다고 할 수 있다.

종단적 학교평가는 두 가지 측면에서 나누어 볼 수 있다. 하나는 흔히 패널자료(panel data)를 이용하여 개별 학생의 학업성취를 시간 혹은 학년에 따라 반복적으로 수집하는 것이다. 이러한 설계를 종단연구설계(longitudinal study design)라고 한다(Linn, 2001). 아래 〔그림 1〕에서는 학생 A의 3학년, 4학년, 5학년, 6학년 때의 성취도가 측정하여 수집된다. 이 설계의 가장 큰 특징은 "동일학생"을 여러 시점에서 반복적으로 측정한다는 것에 있다.

〔그림 1〕 성장평가를 위한 종단자료구조

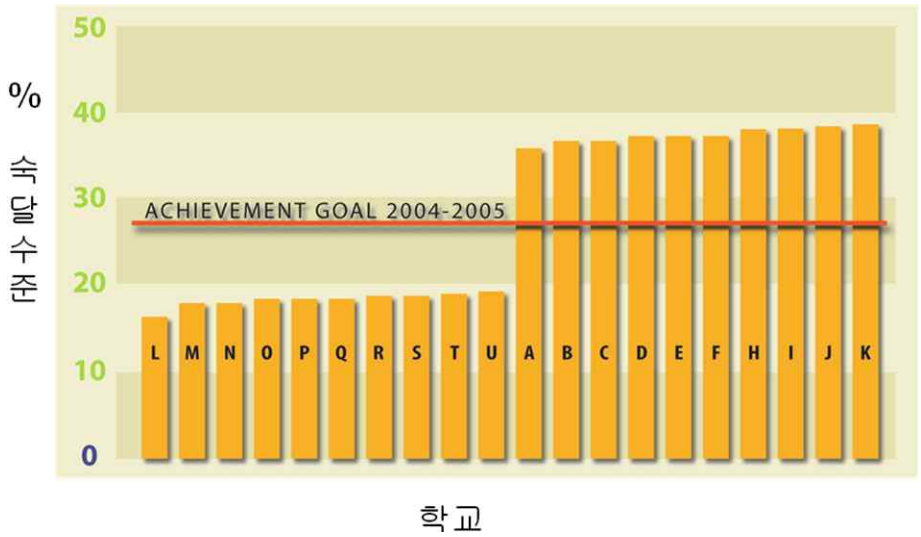
또 다른 종단적 평가방법은 유사종단설계(quasi-longitudinal design)를 이용하는 것이다. 이 방법은 특정 학년 학생의 학업성취도를 매년 측정하여 수집하는 형식을 취한다. 현재의 국가수준 학업성취도 평가 연구의 표집학년인 6학년을 예로 든다면, 2006학년도의 6학년, 2007학년도의 6학년, 2008학년도의 6학년, 2009학년도의 6학년, 2010학년도의 6학년 학생들의 성취도 자료를 수집하여 학교효과 및 교육효과를 종단적으로 평가하는 것이다. 동일 학생의 능력을 반복 측정하는 종단연구설계와 달리, 특정 학년의 성취도를 반복적으로 수년에 걸쳐

수집하기 때문에 유사종단연구설계 혹은 중다 코호트(multiple-cohorts)설계라고도 한다. 이러한 자료구조는 현재 한국교육과정평가원에서 매년 실시하고 있는 국가수준 학업성취도 평가 연구의 연구설계와 동일한 것이다.

종단적 패널자료를 이용한 통계적 분석 모형은 1980년대 이후 성장모형(growth model)혹은 잠재변인 성장곡선 모형(latent growth curve analysis)이란 이름으로 여러 분야에서 많은 연구가 이루어졌다. 이 방법은 학생의 발달적 측면에서 성장곡선의 형태를 추정하고 추정된 성장곡선이 학생의 배경 변인 및 학교변인에 따라 어떻게 다른지를 탐색하는 것을 주된 목적으로 한다(Bryk & Raudenbush, 1987; McGardle & Epstein, 1987; Gibbons, Hedeker, Waternaux, & Davis, 1988; Muthen & Curran, 1997; Choi, 1998; Seltzer, Choi, & Thum, 2003).

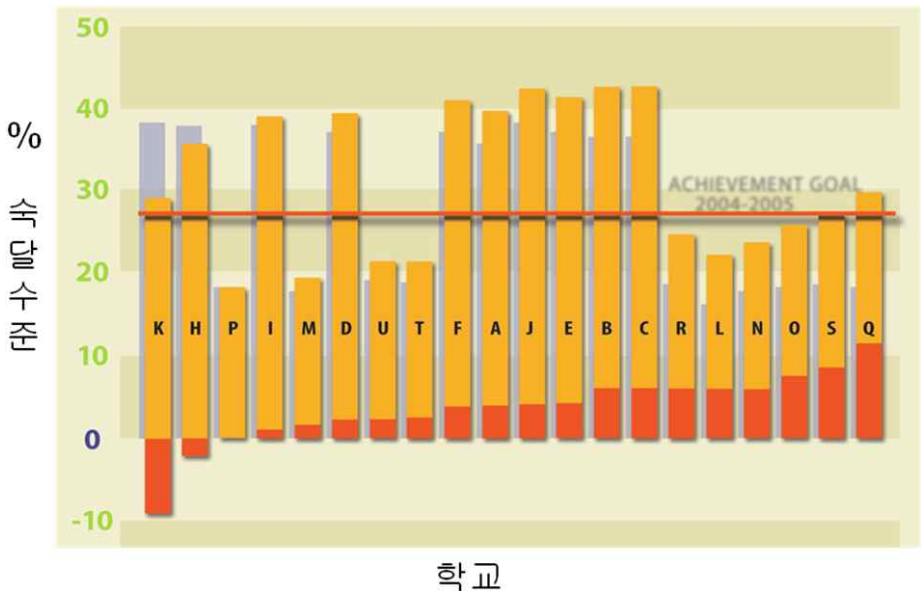
반면, 유사종단연구설계를 이용한 통계적인 모형은 상대적으로 많은 연구가 진행되지 않은 상태이다. 그러나, NCLB 초기에 개별 학생의 성장보다는 학년전체 혹은 학교전체의 연도별 달성목표(Annual measurable objectives) 달성 여부를 판단하는 교육책무성 체계에 적합한 모형으로 인식되어 개선모형(improvement model; Goldschmidt & Choi, 2007; Goldschmidt, Roschenwski, Choi, Auty, Blank, & Williams, 2006)이란 이름으로 모형들이 개발되었다(Raudenbush & Williams, 1995; 최길찬, 2005). 이러한 통계적 모형에 대한 자세한 설명은 다음 절에서 제시한다.

그럼, 현시점 평가와 성장평가의 기본적인 개념을 아래에 제시된 2개의 그림을 기초로 살펴보자. 학생의 학업성취도를 기초로 학교의 교육책무성을 평가하는 가장 전형적이고 쉬운 방법은 매년 달성할 목표를 숙달수준(proficient level)의 관점에서 정하고, 각 학교의 숙달수준에 도달한 학생의 퍼센트가 해당 연도의 달성목표에 도달했는지 확인하는 일일 것이다. 이러한 방법이 현재 시행하고 있는 미국의 NCLB정책이다. 〔그림 2〕는 현시점 평가의 한 예로서 20개의 학교(학교 L 부터 학교 K)가 2004-2005 학년도 숙달수준 달성 목표인 28%를 달성했는지 여부를 보여준다. 각 학교의 숙달수준 %는 학생의 학업성취도 검사의 원점수를 보통의 경우 5개 등급으로(수월수준(advanced), 숙달수준(proficient), 보통(basic), 보통이하(below basic), 최저보통이하(far below basic)) 나누고, 두 번째 상위등급인 숙달수준 등급 이상의 학생의 퍼센트를 계산한 것이다. 예를 들어 학교 L부터 학교 U는 숙달수준 이상의 등급을 받은 학생의 퍼센트가 20% 미만이며, 이는 2004-2005학년도에 달성해야 할 28% 숙달수준 목표에 도달하지 못한 학교들이다. 이와는 반대로, 학교 A부터 학교 K 까지는 숙달수준 목표를 모두 달성한 학교들이다.



〔그림 2〕 현시점 평가: 숙달수준과 년도별 달성목표

* 재인용(Goldschmidt & Choi, 2007)



〔그림 3〕 성장 평가: 2년간 숙달수준의 변화

* 재인용(Goldschmidt & Choi, 2007)

이러한 현시점 평가를 2년에 걸쳐 실행한다고 가정해 보자. 〔그림 2〕에 제시된 것처럼 2004-2005학년도의 목표 숙달수준을 각 학교가 달성했는지 여부로 판단하는 것처럼

2005-2006학년도 역시 같은 방식으로 현시점 평가를 수행할 수 있다. 12개의 학교(학교 K, H, I, D, F, A, J, E, B, C, Q, S)는 2005-2006학년도 목표 숙달수준을 달성하였고, 나머지 10개 학교는 목표를 달성하지 못한 것으로 나타났다. 2004-2005년과 비교해서 볼 때, 2004-2005년도에 목표를 달성했던 10개 학교 이외에 학교 Q와 S가 추가적으로 2005-2006년 목표를 달성한 것으로 나타났다.

다음으로 성장의 관점을 적용해서 학교수행을 살펴보자. 학교 K는 2년 연속으로 해당년도의 숙달목표를 달성하였다. 하지만, 전년도에 비해 이번 연도에는 숙달수준에 있어 10%정도의 감소를 보였다. 이에 반해, 학교 O는 2년 연속으로 숙달목표를 달성하지 못했지만, 2005-2006학년도에는 전년에 비해 약 7-8% 정도의 증가를 보였다. 이러한 경우에 학교 K와 학교 O 중에서 어떤 학교가 더 보상을 받아야 하는가?

또 다른 두 그룹의 학교들을 비교해 보자. 학교 F, A, J, E, B, C는 2년 연속으로 해당년도 숙달목표를 달성했을 뿐 아니라 전년도 대비 2005-2006학년도 숙달수준도 증가하였다. 또 다른 그룹의 학교는 학교 R, L, N, O 이다. 이들 학교는 2년 연속으로 목표를 달성하지 못했지만, 성장의 관점에선 성공적인 학교라고 할 수 있다. 보통의 경우 목표를 달성하지 못한 학교들이 교육여건이나 학생들의 전반적인 사회경제적 지위가 다른 학교들에 비해 낮다고 가정할 때, 이러한 학교들의 성장은 어려운 여건을 극복하고(breaking odds) 이룬 보다 가치로운 결과라고 판단할 수도 있다.

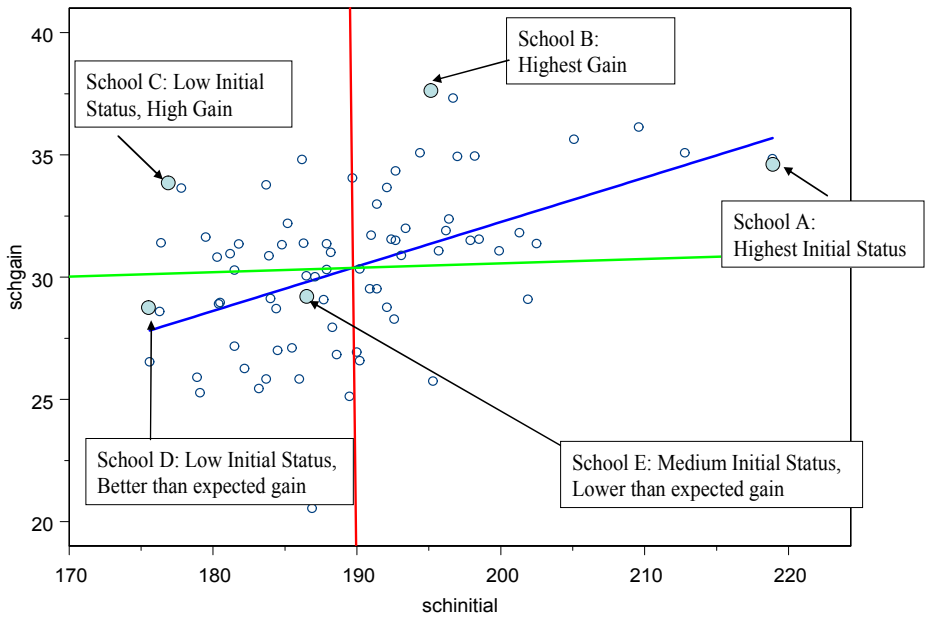
이와 같이 현시점 평가와 성장평가는 학교수행에 대한 다른 정보를 제공한다. 이는 “어떤 학교가 더 효과적인 학교이고, 그에 적합한 평가를 받아야 하는 학교인가?”라는 질문에 대한 대답이 관점에 따라 달라질 수 있음을 의미한다.

이러한 현시점 평가와 성장평가의 결과가 다를 수 있다는 것을 염두에 두고 논의를 한 단계 진전시켜보자. 현시점 평가나 성장평가를 하는데 있어서 고려해야 문제가 있다. 첫 번째는 학생의 성취도 점수를 어떻게 학교수준으로 통합해서 나타내느냐 하는 것이고, 두 번째는 어떻게 하면 학교의 평가가 공정한 룰에 의해서 이루어질 수 있는가 하는 것이다. 첫 번째 질문은 검사척도의 성격과 연관되어 있는 문제이다. 먼저, 앞의 예시에서 보는 바와 같이 개별 학생의 점수를 등급화하고, 각 등급의 학교 퍼센트를 계산하는 방식이 있을 수 있다. 이는 개별 정보를 유목화 하여 나타내고, 이해가 간편하며, 개별학생 한 명 한 명의 정보가 학교 수행수준 평가에 잘 반영된다는 장점이 있다. 하지만, 어떻게 하면 수행수준을 의미있게 등급화 할 수 있을 것인가 하는 문제가 남아 있고, 퍼센트로 나타난 학교수행 수준은 종단적으로 통계적 모형을 통해 분석하기에 적절치 않은 척도라는 단점이 있다.

이와는 달리 학생의 성취도 점수의 학교평균을 기초로 학교평가를 수행하는 방법도 있다. 이는 가장 단순하게는 산술적으로 학생의 학업성취도 점수의 학교평균을 계산하여 학교 간 비교를 하는 방법이 있을 수 있고, 혹은 다층자료분석(multilevel analysis) 방법 등의 다양한

통계적 모형을 적용하는 방법이 있을 수도 있다. 이에 대해서는 다음 절에서 자세히 설명하기로 한다. 다만, 이 방법의 한 가지 단점 혹은 유의점은 평균치는 극단의 점수(outliers)의 영향을 많이 받기 때문에, 소수의 아주 낮은 점수를 받은 학생 혹은 아주 높은 점수를 받은 학생 때문에 학교평균치가 학교전체 대표값으로 적절치 못한 경우가 발생한다는 것이다. 이러한 극단값의 영향력을 통계적으로 고려하는 방법들이 제시되어 있다(Seltzer, Wong, & Bryk, 1996; Seltzer, Novak, Cho, & Lim, 2001; Seltzer & Choi, 2003).

두 번째 문제로 학교 간 평가를 위한 평가준거의 문제에 대해 살펴보자. 앞의 예에서 현시점 평가와 성장평가가 학교효과나 학교수행에 대한 다른 정보를 제공한다는 것을 보여주었다. 이는 기본적으로 평가준거가 다르기 때문이며, 이러한 평가준거로 중요하게 고려되는 것이 "비슷한 것끼리의 비교(comparison like with like)"의 개념이다.



〔그림 4〕 다른 평가준거에 따른 학교효과 순위의 변동

1) 초기상태(X축): A > B > E > C > D
2) 성장(Y축): B > C > A > E > D
3) 초기상태를 통제한 후의 성장(회귀선): B > C > D > E > A
학교의 초기상태 학교전체 평균은 190이며, 학교성장 전체평균은 평균은 30임.
* 재인용(Choi, Goldschmidt, & Yamashiro, 2005)

개인이나 학교는 서로 다른 특성을 가지고 있으며, 그런 특성 중에서 어떤 것들은 개인

혹은 학교의 통제범위 밖의 것들이다. 예를 들어, 학생의 경우에 있어서 부모의 사회경제적 지위는 학생의 학업성취도에 영향을 주는 중요한 변인임에는 틀림없지만, 이는 학생에게 "주어진" 조건이다. 학교의 경우에 있어서도 학교에 입학한 학생들의 사회경제적 지위나 학생의 입학당시의 학력수준은 학교에게 주어진 조건이라고 할 수 있다. 학생 혹은 학교를 평가함에 있어서 이러한 어쩔 수 없는 조건들을 통계적으로 통제하고 유사한 조건을 가진 학생이나 학교들 간에 비교를 한다면 보다 공정한 평가가 이루어질 수 있다는 주장들이 있다.

〔그림 4〕는 평가준거에 따라 학교효과의 학교 간 순위가 달라지는 것을 보여준다. X축은 학교의 전년도 혹은 초기상태의 학교평균이며, Y축은 2년간의 획득점수(gain score)의 학교평균이다. 그리고, 이 두 변인 간의 관계를 나타내는 회귀선이 제시되어 있다. 먼저, 현시점 평가의 준거에서 평가한다면, 초기상태(전년도 학교평균, X축)의 크기에 따라 학교순위를 정할 수 있다. 이 경우, 예시로 제시된 5개의 학교의 순위는 학교 A, B, E, C, D의 순서이다.

반면, 성장평가의 관점에서 학교의 순위는 Y축상의 크기에 따라 결정되며, 이때의 학교순위는 B, C, A, E, D가 된다. 이처럼 두 평가준거에 따른 학교순위는 많이 다르며, 이는 〔그림 2〕와 〔그림 3〕에서 제시한 것과 유사하다. 학교 A의 경우, 전년도의 학교평균은 5개 학교 중 가장 높았지만, 2년 간의 성장에 있어서는 3위에 해당하였으며, 학교 C의 경우 전년도에는 5개 학교 중 학교평균이 4위였지만, 성장의 관점에서는 2위이다.

회귀선을 이용하여 추가적인 평가준거를 적용해 볼 수도 있다. 회귀선은 X축의 변인과 Y축의 변인관계를 나타내며, 회귀선은 각 X값에 대한 Y값의 분포의 평균값을 찾아 연결한 선이다. 따라서 이 회귀선을 평가준거로 적용한다는 의미는 비슷한 조건을 가진 학교들(즉, X축값, 초기상태)을 획득점수 평균(Y축값)에 따라 평가한다. 예를 들어, 학교 C와 D를 비교해 보자. 이 두 학교의 초기상태(X축값)는 약 176점 정도로 비슷하다. 하지만 획득점수 평균(Y축값)는 학교 C가 D보다 약 5점 정도가 높다. 그러나 이 두 학교는 비슷한 초기 상태를 가진 학교들보다 획득점수가 높은 학교이다. 왜냐하면, 이 두 학교의 Y값이 회귀선 위에 놓여 있기 때문이다. 통계적으로 말하자면, 이 두 학교의 잔차(residual: 회귀선에 의한 기대값과 실제 관찰값과의 차)는 모두 양의 값을 갖는다.

이 잔차화된 획득점수(residualized gain)를 평가준거로 위의 5개 학교의 순위를 정하면, B, C, D, E, A의 순이다. 이들 중에 학교 E와 A는 회귀선 아래에 위치하기 때문에 그들의 잔차는 마이너스이며, 이는 초기상태가 비슷한 다른 학교들에 비해 획득점수 평균이 낮다는 것을 의미한다. 이처럼 어떤 조건을 학교 간에 통계적으로 동일화해서 평가했을 경우, 평가결과는 많이 달라지게 된다.

Ⅴ. 성장평가를 위한 통계모형

1. 개인 성장모형

개인 성장모형은 개인의 반복적 측정치를 이용한 종단연구설계에 이용되는 통계적 모형이다. 아래 수식 (1)에서 (4)에 걸쳐 제시된 모형은 3수준 위계모형이다. 이 모형은 학생의 반복 측정치가 학생에 내재되고, 학생은 학교에 내재되는 자료구조에 부합하는 모형이다. 1수준(학생 내 모형)에서 종속변인인, 이 학교 j에 속한 학생 i의 t 시점에서의 학업성취도 점수(Ytij)는 시간변인(atij)의 함수로 나타난다. 1수준 모형은 개별 학생의 성장곡선의 형태를 결정하게 되는데, 반복측정치 자료의 수 및 반복측정치의 관찰치 성장곡선의 형태에 따라 성장곡선의 형태를 모형화한다. 가장 단순한 형태는 선형(linear)이지만, 2차곡선(quadratic), 3차곡선(cubic)등의 다양한 형태를 모형화할 수 있다. 설명의 편의상 수식(1)에 제시된 모형은 성장곡선이 선형이라고 가정하고, 3학년부터 6학년까지 4번의 학업성취도 점수를 수집했다는 가정하에, 시간변인(a_{tij})은 0, 1, 2, 3 으로 코딩했다고 하자. 시간변인을 이런 방식으로 코딩하면, 모형의 절편인 π_{0ij}는 학교 j 에 속한 학생 i의 3학년때의 성적을 나타내고(종단자료의 첫 번째 시점의 상태를 의미하기 때문에 흔히 초기상태라고 함), 모형의 기울기(π_{1ij})는 그 학생의 학년 간 평균 변화율을 나타낸다. 참고로, 시간변인을 -3, -2, -1, 0으로 코딩하면, 기울기는 여전히 학생의 학년 간 평균 변화율을 나타내고, 절편은 6학년 때의 상태를 의미한다(단, 이 때 절편은 초기상태가 아니다). 초등학교의 마지막 학년인 6학년 때의 성적은 초등학교 전체의 학교효과를 반영하는 점수라고 가정할 때, 이러한 코딩 방법도 연구문제에 따라 잘 활용될 수 있다.

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} a_{tij} + \epsilon_{tij} \qquad \epsilon_{tij} \sim N(0,\sigma^2) \tag{1}$$

아래의 2수준(학생 간; 학교 내)모형은 학교의 평균 초기상태와 평균 변화율을 추정한다. 수식 (2a)의 β_{00j}는 학교 j의 평균 초기상태(3학년 성적 평균)를 나타내고, 수식 (2b)의 β_{10j}는 학교 j의 평균 변화율을 나타낸다. 흔히 무선효과라 불리는 r_{0ij}와 r_{1ij}는 각각 학생 i의 초기상태와 변화율과 그 학생이 속한 학교 평균 초기상태와 평균변화율의 편차를 의미한다. 이들의 평균이 0 이고, 변량이 $\tau_{\pi 0}$와 $\tau_{\pi 1}$인 정상분포를 따른다고 가정된다.

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \qquad r_{0ij} \sim N(0,\tau_{\pi 0}) \tag{2a}$$

$$\pi_{1ij} = \beta_{10j} + r_{1ij} \qquad r_{1ij} \sim N(0,\tau_{\pi 1}),\ Cov(r_{0ij},r_{1ij}) = \tau_{\pi 0\pi 1} \tag{2b}$$

위의 모형에는 학생개인배경 변인이 모형내에 포함되지 않았다. 앞서 언급한 바와 같이 개인배경변인을 모형에 포함시켜서, 이들 변인의 차이를 통계적으로 통제한 후의 학교 초기상태와 평균변화율을 산출할 수 있다.

$$\beta_{00j} = \gamma_{000} + u_{00j} \qquad u_{00j} \sim N(0,\tau_{\beta 00}) \tag{3a}$$

$$\beta_{10j} = \gamma_{100} + u_{10j} \qquad u_{10j} \sim N(0,\tau_{\beta 10}),\ Cov(u_{0ij},u_{1ij}) = \tau_{\beta 00\beta 10} \tag{3b}$$

마지막으로 위의 수식 (3a)와 (3b)는 3수준(학교 간) 모형을 나타낸다. γ_{000}와 γ_{100}는 각각 표집된 전체 학교의 평균 초기상태와 평균 변화율을 나타낸다. 무선효과인 u_{00j}와 u_{10j}는 각각 학교 j의 전체 학교 평균 초기상태로부터의 편차, 그리고 전체 학교 평균변화율로부터의 편차를 나타낸다. 따라서 학교 A의 평균 초기상태가 표집된 전체 학교의 평균 초기상태보다 높으면, 학교 A의 u_{00j}값은 양의 값을 갖게 된다. 일반적으로 이러한 모형을 통해 학교평가를 수행할 때 학교의 순위는 이러한 각 학교별 무선효과의 크기에 따라 결정된다.

2. 코호트 성장모형

개인 성장모형과 코호트 성장모형의 가장 큰 차이점은 같은 학생을 반복적으로 측정한 자료를 이용하느냐 아니면 같은 학년의 학생을 반복적으로 측정하느냐에 있다. 수식 (4)은 1수준(코호트내)모형이다. 〔그림 1〕에 제시한 것처럼 2006년도부터 2010년까지의 6학년의 학업성취도 자료를 분석한다고 가정해보자. 이 때 각 학교는 5년간(5개의 코호트)의 6학년 학생의 학업성취도 성적을 갖게 된다. 이 때 최대의 관심사는 “5년간의 6학년 학업성취도가 어떻게 변화하는가” 하는 질문과 “학교 간의 변화성장률에 차이가 있는가” 일 것이다.

$$Y_{ijk} = \beta_{0jk} + r_{ijk} \qquad r_{ijk} \sim N(0,\sigma^2) \tag{4}$$

종속변인 Y_{ijk}는 학교 k, 코호트 혹은 연도 j에 속한 학생 i의 학업성취도 점수이다. 수식 (4)에서 β_{0jk}는 학교 k, 연도 j에 속한 학생들의 평균이다. 즉, 1수준 모형은 학생이 속한 연도 즉 코호트의 기대값(평균값)을 추정하는 것이 주 목적이다.

아래의 2수준(코호트간; 학교 내)모형은 1수준 모형에서 추정된 평균값을 이용하여 성장모형을 적용한다. 즉, 시간변인인 $year_{jk}$는 처음 연도를 0으로 코딩하고, 6학년 평균성적이 선형적으로 증가한다는 가정하에 나머지 연도를 각각 1, 2, 3, 4로 코딩할 수 있다. 이 때, θ_{0k}는 학교 k의 2006년도 6학년 학생의 평균을 나타내는 모수치이며, 이것은 첫 번째 측정년도

의 평균 모수치이기 때문에 학교 k의 초기상태라 할 수 있다. θ_{1k}는 학교 k의 6학년 학생 평균의 5년 간의 평균 변화율을 나타내는 모수치이다.

$$\beta_{0jk} = \theta_{0k} + \theta_{1k} year_{jk} + u_{0jk} \qquad u_{0jk} \sim N(0, \tau_{\beta 0}) \qquad (5)$$

설명의 편의상 성장곡선의 모양이 선형을 나타내도록 시간변인을 코딩하였지만, 관찰된 성장곡선의 모양을 검토한 후 다양한 모양을 성장곡선을 추정할 수 있으며, 혹은 분할된 시점들간에 다른 형태의 성장곡선을 모형화할 수 있다(e.g., interrupted time-series model, piecewise growth model).

아래의 수식 (6a)와 (6b)는 3수준(학교 간) 모형을 나타낸다. Φ_{00}는 초기상태의 전체 표집학교의 평균이고, Φ_{10}는 전체 표집학교의 평균변화율이다. 앞에 제시된 개인 성장모형의 경우와 유사하게 V_{0k}와 V_{1k}의 크기에 따라 개별 학교 간의 비교를 수행한다.

$$\theta_{0k} = \Phi_{00} + V_{0k} \qquad V_{0k} \sim N(0, \tau_{\theta 0}) \qquad (6a)$$

$$\theta_{1k} = \Phi_{10} + V_{1k} \qquad V_{1k} \sim N(0, \tau_{\theta 1}), \quad Cov(V_{0k}, V_{1k}) = \tau_{\theta 0 \theta 1} \qquad (6b)$$

그런데, 위에 제시된 두 모형에서 학교평균 변화율은 초기상태를 통제하지 않은 모형에 의해 추정된 것이다. 초기상태를 통제한 후의 변화율을 추정하기 위해서는 초기상태를 변화율의 독립변인으로 모형 내에 포함해야 한다. 구체적으로 말하자면, 수식 (2a)의 학생초기상태(π_{0ij})가 수식 (2b)에 독립변인으로 포함되어야 하며, 학교수준 모형에서도 수식 (3a)의 학교평균 초기상태(β_{00j})는 학교 평균변화율을 설명하는 독립변인으로 수식 (3b)에 포함되어야 한다. 코호트 성장모형에서도 마찬가지로 수식 (6a)의 학교평균 초기상태(θ_{0k})는 수식 (6b)에 포함되어야 한다(즉, $\theta_{1k} = \Phi_{10} + \theta_{0k} \times b + V_{1k}$). 이러한 통계적 모형은 위계모형(hierarchical model)내에 잠재변인(latent variable)들 간의 회귀모형을 통합한 것이며, 최근에 이에 대한 연구가 활발히 진행되고 있다(Seltzer, Choi, & Thum, 2003; Klein & Muthen, 2006; Harring, 2009; Choi & Seltzer, 2010).

현행 국가수준 학업성취도 평가는 앞서 〔그림 1〕의 설명에서 제시한 바와 같이 동일 학년의 학생을 반복적으로 측정하는 유사종단설계 혹은 코호트 설계에 따른 평가방법이다. 올해가 초등학교 6학년, 중학교 3학년, 고등학교 1학년 전집을 대상으로 평가를 실시한 2년째 되는 해이기 때문에, 4년째가 되는 2년 후에는 첫 번째 해의 초등학교 6학년이 중학교 3학년이 되며, 그 후 1년 후에는 고등학교 1학년이 되고, 다시 2년 후에는 고등학교 3학년이 되며 대학수학능력시험을 치르는 연령이 된다. 국가수준 학업성취도 평가와 대학수학능력시험에

대한 국가적인 데이터베이스가 구축되어 있다면, 이론적으로 한 학생의 학업성취도의 변화를 종단적으로 추정할 수 있으며, 이러한 학생의 학업성취도의 변화에 학교급별로(초등학교, 중학교, 고등학교) 어떤 영향을 주었는지를 통계적으로 모형을 설계할 수도 있다. 뿐만 아니라, 2년 전에 국가수준 학업성취도 평가에 참여한 초등학교 6학년 학생을 첫 번째 코호트라고 하고, 올해 초등학교 6학년을 두 번째 코호트, 그리고 세 번째, 네 번째 코호트의 학생들의 자료를 축적해 간다면, 학업성취도 변화의 코호트 간 차이도 역시 통계적으로 추정이 가능하다. 이러한 모형은 위에서 제시한 개인 성장모형과 코호트 성장모형을 통합한 중다코호트 개인 성장모형(multiple-cohort longitudinal growth model: Choi, 2009)으로 불린다.

Ⅵ. 맺음말: 몇가지 정책 제언

이 글은 2년째 수행한 국가수준 학업성취도 평가연구를 통해 얻어진 자료를 활용한 학교평가의 방향, 평가 설계, 평가 지표에 대한 이해, 검사척도, 통계적 모형에 대해 종합적으로 고찰하였다. 이러한 논의를 기초로 다음의 몇 가지 정책적 제안을 제시한다.

첫째, 국가수준 학업성취도 평가에 대한 의사결정을 전문적으로 자문하는 기구의 설치가 필요하다. 국가수준 학업성취도 평가연구는 국가적인 규모로 학생의 학력을 측정하고, 이를 교육적으로 활용하는 국가적 사업이다. 이러한 국가적 사업을 통해 수집된 자료가 국가수준의 교육에 관한 중요한 정보가 되고, 이를 교육정책의 평가 및 새로운 정책 입안을 위한 기초 자료로 활용하기 위해서는 보다 체계적이고 구체적인 청사진이 필요하다. 여기에는 평가문항 개발을 위한 내용 영역 설정, 문항 유형, 검사 간 동등화 등 검사와 문항에 대한 세부적이고 기술적인 계획에서부터, 수집한 자료를 어떻게 데이터베이스화 하고, 결과를 어떻게 지역 교육청, 학교, 교사, 학생, 학부모에게 제시하는가 하는 결과 보고(reporting)의 문제, 어떤 통계적 모형을 이용하여 분석하는가 하는 통계모형 설정의 문제, 분석 결과를 어떻게 학교평가 및 교육정책 평가 혹은 새로운 교육정책 입안을 위해 활용하는가 하는 문제 및 교육지원 및 모니터링 시스템에 대한 청사진 등 아주 다양하고 광범위한 계획이 포함될 수 있다. 현재 이러한 부분에 대한 의사결정은 관련당사자들의 의견과 여론을 수렴하면서 교육과학기술부와 한국교육과정평가원을 중심으로 이루어지고 있다. 그런데, 국가수준 학업성취도평가에 대한 갈등하는 다양한 의견들이 제시되고 있는 만큼 교육전문가, 정책결정자, 지역 교육청, 교장 및 교사, 학부모, 연구기관 등으로 구성된 '국가교육평가위원회' 같은 기구를 구성하여 의견과 여론을 수렴하여 이해관계 당사자들간의 의견대립을 해소하면서 국가적 수준의 정책결정에 도움을 주도록 해야 한다.

둘째, 국가적 수준의 교육정보 시스템을 조속하게 구축하는 것이 필요하다. 국가수준 학업성취도 평가 자료, 대학수학능력시험 자료, 고등학교 3학년의 모의 고사 자료 등은 한국교육과정평가원에서 관리하고 있다. 그러나, 그러한 교육정보들이 다양한 변인들과 연계하여 국가적 수준에서 체계적인 데이터베이스로 구축되어 있다고 말하기 어렵다. 시스템 구축을 위해선 가장 기초적인 학생, 교사, 학교에 대한 고유 아이디 부여로부터, 학업성취도뿐 아니라 학생의 배경변인, 정서 및 사회적 요인 변인, 교사 배경 변인, 학교 배경 변인 등 다양한 정보들이 포함될 수 있으면, 이러한 정보가 수년에 걸쳐 자료로 축적되었을 때 그 가치는 생각하는 것 이상으로 매우 크다고 할 수 있다. 뿐만 아니라, 한국교육개발원과 한국직업능력개발원 등에서 수집한 종단자료들이 함께 연계되어 활용된다면 그 효과는 더 커지리라 생각된다. 따라서 한국교육과정평가원과 한국교육학술정보원 등이 협력하여 성취도 정보들을 국가적 수준의 데이터베이스로 조속하게 구축해야 한다.

셋째, 대학수학능력시험을 통한 국가수준 학력의 변화 측정에 대한 논의가 이루어져야 한다. 약 15년에 걸쳐 수행한 대학수학능력시험은 문항의 질에서나 검사관리 및 채점에 있어서도 한국 최고의 학업성취도 자료라고 할 수 있다. 이러한 자료를 잘 분석하고 활용하면 중요한 교육정책 평가 및 국가적 수준의 교육정보로 활용될 수 있다. 대학수학능력 시험 성적이 지난 해부터 공개되고(김성열, 2009), 한국교육과정평가원을 중심으로 부분적인 분석이 이루지고 있다(한국교육과정평가원, 2009). 수능성적 공개의 초기 단계여서 대중적 관심은 지역간·학교간 격차 자체에 관심이 높은 게 사실이다. 물론 분석단위간 격차 여부와 정도를 확인하는 것도 필요하지만, 그 격차의 원인과 변화추세를 확인하는 것이 더욱 필요하다고 할 수 있다. 그리고 15년간 교육과정이 6차, 7차, 2007 개정, 2009 개정 교육과정으로 변천하였는데, 이렇게 교육과정이 바뀜에 따라 국가적 수준에서 보면 학생들의 학력이 점점 높아졌는가 하는 질문을 던질 수도 있다. 또한, 그간 공교육의 근간을 이룬 고교 평준화 정책에 대한 평가도 제한적이긴 하지만 이 자료의 분석을 통해 가능하리라 생각된다.

마지막으로, 교육정책 평가 및 새로운 교육정책 입안을 위한 경험과학적 방법을 이용한 자료분석 방법이 활성화되어야 한다는 점을 강조하고 싶다. 이는 국가적인 교육정책이나 교육사업이 경험적 자료에 근거하고 과학적 연구방법에 기초한 평가결과를 수집하여 활용함을 의미한다. 이명박정부가 들어선 이후에 증거에 기반한 새로운 교육정책 입안이 강조되고 있다. 교육에 관한 정보를 공개하고, 이를 실증적으로 분석한 결과에 기초한 정책결정이 이루어지기 시작한 만큼, 어떤 정보를 왜 수집하고, 어떻게 분석하고, 어떻게 결과를 활용할지에 대한 기본적인 계획 및 모범사례가 정부 및 학계의 협의에 의해 마련되고, 활성화되었으면 하는 바람이다.

참 고 문 헌

김성열(2009). 대학수학능력시험 성적 분석 결과. **대학수학능력시험 성적 분석 결과 전문가 세미나 자료집.** 한국교육과정평가원.

김성열 · 남명호 · 정은영 · 김성숙(2009). **국가경쟁력 제고를 위한 국가수준 학업성취도 평가의 발전방향.** 한국교육과정평가원 포지션페이퍼, ORM 2009-5-1.

최길찬(1998). 성장모형의 두 다른 접근: 위계모형과 구조방정식 모형. 황정규 편. **교육평가의 새지평** (pp. 357-417). 서울: 교육과학사.

최길찬(2005). 유사종단자료를 이용한 학교효과 추정 위계적 모형: 척도점수와 NCE 점수와의 비교. **아시아교육연구, 6**(1), 59-81.

한국교육과정평가원(2009). **수능 및 학업성취도 평가 결과 분석 심포지엄.** 연구자료, ORM 2009-40.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147-158.

Choi, K. (2009). Multisite multiple-cohort growth model with gap parameter(MMCGM): Latent variable regression 4-level hierarchical models. *IES unsolicited grant annual report.* Washington, D.C.

Choi, K., & Seltzer, M. (2010). Modeling heterogeneity in relationships between initial status and rates of change: treating latent variable regression coefficients as random coefficients in a three-level hierarchical model. *Journal of Educational and Behavioral Statistics, 35*(1), 54-91.

Choi, K., Goldschmidt, P., & Yamashiro, K. (2005). Exploring models of school performance: from theory to practice. In J. Herman & E. Haertel (Eds.), Data use and misuse. *The 104th Year book of the National Society for the Study of Education.*

Gibbons, R., Hedeker, D., & Waternaux, C., & Davis, J. (1988). Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin, 24*(3), 438-443.

Goldschmidt, P., & Choi, K. (2007). *The practical benefits of growth models for accountability and the limitations under NCLB.* Policy Brief 9, Spring, 2007. Center for Research on Evaluation, Standards and Student Testing (CRESST). Los Angeles: University of California.

Goldschmidt, P., Roschewski, P, Choi, K., Auty, W., Blank, R., & Williams, A. (2006).

Policymaker's guide to growth models for school accountability: how to accountability model differ? The Council of Chief State School Officers .Washington, D.C.

Harring, J. (2009). A Nonlinear Mixed Effects Model for Latent Variables. *Journal of Educational and Behavioral Statistics, 34*(3), 293-318.

Klein, A., & Muthen, B. (2006). Modeling Heterogeneity of Latent Growth Depending on Initial Status. *Journal of Educational and Behavioral Statistics, 31*(1), 357-375.

McGardle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110-133.

Muthen, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*(4), 371-402.

Raudenbush, S. W., & Willms, D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-401.

Seltzer, M., Wong, W., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21*(2), 131-167.

Seltzer, M., Novak, J., Choi, K., & Lim, N. (2002). Sensitivity analysis for hierarchical models employing t level-1assumptions. *Journal of Educational and Behavioral Statistics, 27,* 181-222.

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*(3), 263-286.

• *논문 접수 : 2010년 5월 1일 / 게재 승인 : 2010년 5월 27일*

ABSTRACT

Monitoring School Performance Based on National-level Achievement Test

Kilchan Choi(Senior Researcher, CRESST/UCLA)
Seong-Yul Kim(President, Korea Institute for Curriculum and Evaluation)

Our purpose in writing this paper is to present and discuss a statistical model that is suitable for the national-level achievement test that has been implemented for the past two years. We first examine the multiple aspects of school performance evaluation - 1) perspectives of monitoring school performance including goals, contents, evaluation units, and status vs. growth; 2) school performance indicators; 3) test metrics such as percentiles, normal curve equivalents, and vertically equated scale score; 4) status model vs. growth model; 5) longitudinal data structure for measuring growth over time; 6) the consequence of evaluating school performance based on different criteria. The principle of design in national-level achievement test is called quasi-longitudinal, or cohort design, and measures the same grade of students over time in contrast to true longitudinal, or panel design, in which the same students are measured over time. In the second part of this paper, we present two different 3-level hierarchical models; one model for the true longitudinal design and the other for cohort design. Lastly, we provide policy-related suggestions regarding how to take full advantage of the results from the national-level achievement test and analyze the implications of doing so.

Key words : national-level achievement test, school accountability evaluation, status model, growth model, cohort-to-cohort growth, multisite multiple-cohort longitudinal growth model