교육과정평가연구
The Journal of Curriculum and Evaluation
2012, Vol. 15, No. 3, pp. 207~230

성취평가제에서 성취도 평정을 위한 분할점수의 적정성 평가 방안

이 상 하(한국교육과정평가원 연구위원)*
최 혁 준(한국교육과정평가원 연구위원)

≪ 요 약 ≫

이 연구의 목적은 현재 중등학교에 적용되고 있는 성취평가제에서 학기말 성취도 평정의 기준이 되는 기준성취율의 적정성을 평가하거나 수정할 수 있는 방안을 제안하는 것이다. 이 방안의 유용성을 확인하기 위해 ○○광역시에 소재하는 중학교 1학년 2개 학급을 선정하였다. 그리고 5개 교과 담당 교사들이 학생들에 대한 교수·학습 경험을 토대로 주관적으로 평정한 성취도와 학기말 평가결과를 토대로 평정한 성취도를 비교하였다. 교사가 주관적으로 평정한 성취도와 학생들의 학기말 성취도의 분류일치도는 카파계수 κ가 0.23~0.77의 범위에 있었고, 불일치 정도에 가중치를 부여한 가중 카파계수 $\kappa_{(w)}$는 0.52~0.88의 범위에 있었다. 또한, 대조집단 수준설정 방법으로 결정한 분할점수가 실제 학기말 성취도 평정의 기준이 되었던 기준성취율보다 평균 1.3~11.5점 낮은 것으로 나타났다. 피험자-중심의 대조집단 수준설정 방법은 현행 성취평가제에서는 기준성취율의 적절성을 평가하는 수단으로 사용하고, 제도개선이 있을 경우에는 사후에 분할점수를 결정하는 방안으로 사용할 것을 제언하였다.

주제어 : 성취평가제, 준거참조평가, 기준성취율, 분할점수, 수준설정, 대조집단 수준설정

Ⅰ. 들어가기

성취평가제는 「중등학교 학사관리 선진화 방안(2011.12.13. 발표)」의 주요 내용 중 하나로서, 현재 정책적으로 추진하고 있는 교수·학습의 변화에 맞도록 학생들에 대한 평가 방법을 변

* 제1저자 및 교신저자, sangha@kice.re.kr

화시키는 것이다. 성취평가제는 학생들 간 서열 중심의 상대평가에서 학생들이 학교교육을 통해 성취해야 할 목표에 도달한 정도를 평가하는 준거참조평가로 전환하는 것을 의미한다. 국가수준에서 성취평가제를 지원하기 위하여 교수·학습과 준거참조평가의 기준으로 활용할 수 있는 교과목별 성취기준·성취수준을 개발하여 보급하였다. 학교에서는 국가수준에서 개발한 교과목별 성취기준·성취수준을 학교상황에 맞도록 수정하거나 자체적으로 만들어 사용하고 있다. 그리고 지필평가와 수행평가를 통해 교과목별 성취기준에 대한 도달 정도를 평가하여 학생들의 학기말 성취도를 평정하고 있다.

중학교 교과와 고등학교 전문교과의 경우에는 2012학년도 1학년 입학생부터 성취평가제를 이미 적용하고 있다. 그러나 고등학교 보통교과의 경우에는 2013년까지 시범학교를 운영한 후에 2014학년도 1학년 입학생부터 성취평가제를 적용할 예정이다(교육과학기술부, 2011). 따라서 2012학년도 특성화고등학교 및 마이스터고등학교 1학년 학생들의 1학기말 교과별 성적은 기존의 석차9등급과 성취도가 혼재되어 학생생활기록부에 기재되어 있다. 즉, 보통교과의 성적은 기존의 석차9등급제로 표기되었으며 전문교과의 경우에는 'A-B-C-D-E'의 성취도로 표기되었다. 한편, 2012학년도 중학교 1학년 학생들의 1학기말 성적의 경우에는 성취도 표기 방식이 교과에 따라 기존의 '수-우-미-양-가'에서 'A-B-C-D-E' 또는 '우수-보통-미흡'에서 'A-B-C'로 변경되었다.

학교현장에서는 학년 초에 학업성적관리규정을 확정하고, 이를 토대로 학생들의 평가 및 성적 등을 처리하도록 되어 있다. 현행 성취평가제의 경우에도 학생들의 학기말 성취도를 구분하는 분할점수[1]를 미리 결정하고 이를 학업성적관리규정에 명시하도록 하고 있다(한국교육과정평가원, 2012a & 2012b). 따라서 학교에서는 이미 설정된 분할점수에 따라 성취도를 평정하는 것이 적절할 수 있도록 지필평가 및 수행평가의 난이도를 적절하게 유지해야하는 어려움이 있다. 학교단위에서 시행되고 있는 평가도구의 경우, 실증적 자료를 토대로 문항의 심리측정학적 정보를 사전에 얻는 것이 불가능하다. 따라서 분할점수를 미리 고정시켜놓고 교사의 경험과 전문성만으로 문항 및 평가도구의 난이도를 조정하는 방법은 한계가 있을 수밖에 없다.

한편, 지필평가와 수행평가가 모두 이루어진 상태에서 학기단위 성취수준을 구분하여 성취도를 평정하는 방법이 측정학적 관점에서 가장 합리적이기는 하지만, 현실적으로 여러 가지 면에서 운영상의 어려움이 있을 수 있다. 구체적인 평가도구와 피험자 특성을 토대로 성취수준을 구분하는 분할점수를 설정하는 것을 수준설정(standard setting)이라고 하는데, 수준설정 방법에 따라 차이가 있지만 일반적으로 다수의 패널들이 장시간 논의를 거쳐 분할점수를 결정하는 과정들이 포함되어 있다. 그런데 학교단위에서 특정교과의 수준설정을 위하여 적절한 규모의 패널을 구성하는 것이 매우 어려울 수 있다. 또한, 학교단위에서 소규모이기는 하지만 수준설정을

1) 성취평가제에서는 이 분할점수를 '기준성취율'이라고 정의하였다.

위한 패널을 구성할 수 있다고 하더라도, 교사 또는 패널들이 학생들의 내신에 직접적으로 영향을 주는 분할점수를 결정할 때 외부의 압력으로부터 자유롭기 어렵다. 교사를 포함한 패널들이 성취수준을 구분하는 분할점수를 낮추어 학생들이 좋은 성적을 얻을 수 있도록 압박을 받을 수 있다. 그리고 일반적인 수준설정 절차는 하나의 평가도구를 토대로 분할점수를 결정하게 되는데, 성취평가제에서는 지필평가 및 수행평가 성적들을 종합하여 성취도를 평정해야 하기 때문에 하나의 평가도구를 사용할 때보다 더 복잡한 과정과 더 많은 시간이 요구될 것이다.

이 연구의 목적은 현재 중등학교에 적용되고 있는 성취평가제에서 학기말 성취도를 평정하는 기준이 되는 기준성취율의 적정성을 평가하거나 수정할 수 있는 방안을 제안하는 것이다. 이 방안의 유용성을 확인하기 위하여, 교사가 교수·학습 경험을 토대로 평정한 성취도와 지필평가 및 수행평가 결과를 토대로 평정한 성취도가 얼마나 일치하는지, 대조집단 수준설정 방법을 사용하여 설정한 분할점수가 성취평가제의 기준성취율과 어느 정도 일치하는지를 분석하였다.

Ⅱ. 선행 연구

시험 점수의 의미를 해석하기 위해 참조하는 틀이 무엇이냐에 따라서 규준참조평가(norm-referenced evaluation)와 준거참조평가(criterion-referenced evaluation)로 구분할 수 있다(Thorndike, 1997). 규준참조평가는 규준집단에서 개별 학생의 상대적인 위치에 따라 학생들을 평가하는 것을 의미하는데, 우리나라에서 시행되고 있는 규준참조평가의 예로는 대학수학능력시험과 고등학교 내신 석차9등급제 등을 들 수 있다. 한편, 준거참조평가는 학생들이 절대적인 기준에 도달한 정도에 따라 학생들의 성적을 산출하는 평가라고 할 수 있는데, 준거참조평가의 예로는 국가수준 학업성취도 평가와 중학교 내신 성적 등을 들 수 있다. 그런데 준거참조평가의 경우 절대적인 기준에 도달했다는 것을 의미하는 시험점수의 기준을 설정하는 것이 필요하다. 예를 들어, 우리나라에서 자동차를 운전하기 위해서는 도로교통법을 비롯한 각종 교통법규와 기본적인 자동차 작동 원리 등에 대해서 어느 정도 소양이 있어야 하는데, 운전면허 필기시험에서 60점 또는 70점 이상이 되어야 그 정도의 소양이 갖추어진 것으로 보고 60점 또는 70점을 합격점수로 결정한 것이라고 할 수 있다. 이처럼 준거참조평가에서 절대적인 기준에 도달한 정도와 시험점수를 체계적이고 합리적으로 연계하는 것을 수준설정이라고 한다. 이하에서는 수준설정의 방법과 절차에 대해서 알아보고, 성취평가제와 수준설정의 관계에 대해서 논의한다.

1. 수준설정 방법과 절차

준거참조평가의 수준설정 절차에서 가장 우선적으로 정의되어야 하는 개념이 수행수준명(performance level label, 이하 PLL)과 수행수준기술(performance level description, 이하 PLD)이다. PLL은 우수학력, 보통학력, 기초학력 등과 같이 특정한 수행수준에 대한 이름을 나타내며, PLD는 특정한 수행수준에 있는 학생들이 무엇을 수행할 수 있는지에 대한 설명이라고 할 수 있다(Cizek, 2006). 수준설정을 좁은 의미로 해석하면, PLD와 평가도구를 토대로 PLL을 구분하는 분할점수(cut score)를 결정하는 것이라고 할 수 있다. PLD에 기술되어 있는 학생들의 수행수준을 측정할 수 있는 평가도구의 특성에 따라 다양한 수준설정 방법이 사용되고 있다.

우리나라 국가수준 학업성취도 평가의 경우에는 우수, 보통, 기초, 기초미달로 학생들의 성취도 수준을 구분하는 수준설정 방법으로 '수정된 앙고프 방법(modified Angoff method)'을 사용한다. 그리고 국가영어능력평가시험의 읽기영역과 듣기영역 시험에서 A, B, C, F 등급을 구분하는 분할점수를 결정하기 위한 수준설정 방법으로 북마크 방법(bookmark method)을 사용한다. 국제학업성취도 평가인 PISA(Programme for International Student Assessment)와 TIMSS(Trends in International Mathematics and Science Study), 미국의 학업성취도 평가인 NAEP(National Assessment of Educational Progress), 미국의 주단위 책무성 평가의 경우에도 방법과 절차에 차이가 있지만 모두 수준설정을 통해 분할점수를 결정하고 있다.

오늘날 대다수의 수준설정 방법은 참가자들로 하여금 평가도구의 개별 문항 또는 과제들의 특성을 충분히 고려해서 가설적 집단이 얼마나 잘 수행할 것인지를 판단하도록 하고 있다(Cizek & Bunch, 2007). 이와 같이 평가도구의 개별 문항 및 과제들을 충분히 고려해서 분할점수를 결정하는 방법을 평가도구-중심 수준설정 방법이라고 한다(Jaeger, 1989). 평가도구-중심 수준설정 방법에는 앙고프 방법(Angoff method), 네델스키 방법(Nedelsky method), 에벨 방법(Ebel method), 북마크 방법 등이 있다. 평가도구-중심 수준설정 방법의 특징을 알아보기 위해, 앙고프 방법을 적용하여 성취수준의 분할점수를 결정하는 절차를 간략히 살펴보면 다음과 같다(Cizek & Bunch, 2007).

첫째, 분할점수를 결정할 자격이 있는 사람들로 패널을 구성한다.

둘째, 패널들은 PLD를 토대로 각 성취수준에 간신히 도달한 최소능력자(minimally competent examinee, 이하 MCE)의 특징에 대해 논의한다.

셋째, 패널들은 각 성취수준의 MCE가 각 시험문항의 정답을 선택할 확률을 독립적으로 판단한다.

넷째, 각각의 패널들이 문항들에 대해 판단한 확률들을 토대로 패널별 성취수준 분할점수들을 결정하고, 이 분할점수들의 평균 또는 중앙값이 성취수준을 구분하는 최종 분할점수가 된다.

평가도구-중심 수준설정 방법에 따라 진행 절차, 정보제공, 판단기준, 반복 횟수 등에 차이가 있다. 그러나 둘째 단계에서 MCE의 특징을 별도로 기술하는 것이 반드시 필요하고, 셋째 단계에서 평가도구의 개별 문항 또는 과제에 대해서 이들의 수행 정도 또는 수행 수준을 각 방법의 기준에 따라 판단하고, 평가도구의 개별 문항 또는 과제에 대한 판단 결과를 종합하여 분할점수를 결정한다는 점에서 공통의 특징을 갖는다고 할 수 있다.

이와는 대조적으로, 평가도구의 개별 문항 또는 과제의 특성보다는 가설적 집단의 특성에 초점이 맞추어지는 방법을 피험자-중심 수준설정 방법이라고 할 수 있다. 피험자-중심의 수준설정 방법으로는 대조집단 방법(contrast groups method)과 경계집단 방법(boardline group method)이 있다.

대조집단 방법은 구체적인 평가도구 또는 분할점수가 알려지지 않은 상태에서, 평가도구를 통해 측정하고자 하는 특성을 토대로 피험자 집단을 숙달과 미숙달 또는 통과와 미통과 등으로 범주화한다. 그리고 각 피험자 집단의 점수분포를 추정하여 두 집단을 구분하는 분할점수를 결정한다. 이 평가도구의 점수분포에서 분할점수는 각 집단의 피험자들이 다른 집단으로 오분류되는 비율(확률)이 가장 작아지는 점수가 되어야 한다. 이와 같은 분할점수를 결정하는 방법에는 두 점수분포의 평균의 중간지점을 선택하는 방법, 두 점수분포의 중앙값의 중간지점을 선택하는 방법, 두 점수분포 곡선이 만나는 지점을 선택하는 방법, 로지스틱 회귀방정식을 사용하는 방법 등이 있다(Cizek & Bunch, 2007). 반면, 경계집단 방법은 대조집단 방법과는 달리 평가도구를 통해 구분하고자 하는 두 개의 집단에 속하는 피험자들을 선택하는 것이 아니라, 두 집단의 경계에 위치하고 있는 피험자들을 선택하고 이들의 점수분포에서 분할점수를 결정하는 방법이다. 패널들은 평가도구를 통해 측정하고자 하는 능력과 경계집단에 속하는 피험자들의 특성에 대해서 충분히 논의한 다음에 경계집단에 속하는 피험자만을 선택하여 경계집단을 구성한다. 경계집단 수준설정 방법의 분할점수는 경계집단 점수분포의 중심경향치(central tendency)인 중앙값, 평균, 최빈값을 사용할 수 있다.

평가도구-중심 수준설정 방법과 피험자-중심 수준설정 방법은 평가도구 또는 피험자 하나만을 고려한다는 것을 의미하지는 않는다. 정확하게 표현한다면, 분할점수를 결정하는 과정에서 평가도구와 피험자 중에 어느 것을 우선적으로 고려하느냐에 따른 구분이라고 할 수 있다. 즉, 평가도구-중심 수준설정 방법은 평가도구에 포함되어 있는 문항 단위에서 각 성취수준에 있는 피험자들의 수행 정도를 판단하여 평가도구 단위에서의 분할점수를 결정하는 방안이고, 피험자-중심 수준설정 방법은 성취수준별 피험자 집단을 우선 구성하고 구체적인 평가도구에서 수행 정도를 비교하여 분할점수를 결정하는 방안이라고 할 수 있다. 그런데, 평가도구-중심 수준설정 방법에서도 문항 단위의 수행 정도를 판단할 때 피험자의 특성을 고려해야 하고, 피험자-중심 수준설정 방법에서도 성취수준별로 피험자 집단을 구성할 때 평가도구의 특성을 고려하지 않을 수 없다. 따라서 피험자 또는 평가도구 특성 하나만을 고려해서 분할점수를 결정하는 수준설정

방법은 없다고 할 수 있다.

2. 성취평가제와 수준설정

현행 성취평가제는 지필평가와 수행평가를 토대로 개별 학생이 교육과정의 성취기준을 어느 정도 달성하였는지를 측정하여 성취도를 평정하도록 되어 있다는 점에서 준거참조평가의 성격을 가진 학생평가 체제라고 할 수 있다. 성취기준에 도달한 정도를 측정하고 학생들의 성취수준을 구분하여 성취도를 평정하는 절차를 살펴보면, 수준설정과 같은 체계적인 접근법보다는 학생들에 대한 교사의 이해, 교사의 교수학습 경험, 교수학습 내용에 대한 교사의 전문성, 지필평가 및 수행평가 출제 권한 등을 토대로 하는 교사의 주관적인 판단에 전적으로 의지하고 있는 실정이다. 그러나 준거참조평가에서 수행수준 또는 성취수준을 구분하는 분할점수의 타당도를 확보하기 위해서는 분할점수를 판단하는 절차의 이론적 토대와 과정이 체계적이고 타당해야 한다는 제언이 있다(AERA, APA, & NCME, 1999; Hambleton & Pitoniak, 2006).

성취평가제에서는 교과별로 내용기준과 수행기준을 통합한 형태를 취하고 있는 성취기준을 토대로 교수·학습과 평가가 이루어지도록 하고 있다. 또한, 성취평가제에서는 일반적인 수준설정에서 사용하는 PLL과 PLD에 해당하는 것을 찾아볼 수 있다. 즉, 교과에 따라 학기말에 학생들의 성취도를 3개(A, B, C) 또는 5개(A, B, C, D, E)의 수준으로 구분하여 성취도를 평정하도록 되어 있는데, A, B, C 등과 같이 문자로 표기되는 성취도는 평가를 통해 구분하고자 하는 학생들의 성취수준을 나타내는 이름이기 때문에 PLL에 해당한다고 볼 수 있다. 그리고 학기말에 각 성취도를 획득한 학생들의 행동특성을 기술한 '학기단위 성취수준기술'이 있는데, 이것은 수준설정에서 각 성취수준에 있는 학생들이 무엇을 할 수 있는지를 설명하는 PLD에 해당한다고 할 수 있다. 그러나 학교단위에서 평가도구를 토대로 성취수준 분할점수를 결정하는 일반적인 수준설정 절차를 바로 적용할 수 없는 현실적인 문제점들이 있다.

우선, 중학교와 특성화고등학교 및 마이스터고등학교에서 성취평가제를 운영하는 절차를 살펴보면, 학기 초에 학업성적관리규정에 지필평가와 수행평가를 토대로 학기말에 성취도를 평정하는 기준점수에 해당하는 '기준성취율'을 미리 규정하도록 되어 있다. 즉, 학교단위에서 수준설정의 최종 목표가 되는 분할점수를 미리 결정한 다음에, 교과별 성취기준들과 학기단위 성취수준기술을 마련하도록 되어 있다. 이것은 성취기준과 학기단위 성취수준기술을 충분히 고려하여 지필평가 문항과 수행평가 과제를 선정하고, 기준성취율에 따라 학기말 성취도를 평정하는 것이 적절하도록 지필평가와 수행평가의 전반적인 난이도를 조절해야 한다는 것을 의미한다. 즉, 수준설정의 관점에서 보면 평가도구와 피험자 집단의 특성을 고려하지 않은 채 성취도를 구분하는 분할점수를 미리 정해놓고, 그 분할점수에 따라 학생들의 성취도 수준이 적절하게 구분될 수 있

도록 지필평가 문항과 수행평가 과제를 출제해야 한다는 것이다. 평가도구의 문항 또는 과제가 구체적으로 주어진 상태에서 PLD를 고려하여 분할점수를 결정하는 일반적인 수준설정 절차에서도, 수준설정 방법에 따라 분할점수에 큰 차이가 있었을 뿐만 아니라 동일한 수준설정 방법을 사용했더라도 패널의 특성에 따라 분할점수에 큰 차이가 발생한다는 연구 결과들이 있다(Jaeger, 1989). 이것은 성취수준을 구분하는 분할점수를 미리 고정시켜놓고 그러한 결과가 나오도록 지필평가와 수행평가를 출제하는 것은 매우 어려운 과제일 수 있다는 점을 시사하고 있다.

또한, 학업성적관리규정이 개정되어 학기말에 성취수준의 분할점수를 결정하는 것이 허용된다고 하더라도, 학교단위에서 일반적으로 많이 사용되고 있는 평가도구-중심 수준설정 방법을 적용하는 것은 매우 어렵다. 전통적으로 교육평가 프로그램의 수준설정에 참여하는 패널은 대표성 있는 20~30명 정도의 사람으로 구성하는 것이 이상적이라고도 하고(Morgan & Michaelides, 2005), 연구결과를 토대로 학업성취도 평가에서 분할점수를 설정하기 위해서 필요한 패널의 수를 10~15명으로 제안하기도 하며(Zieky & Perie, 2006), SAT 추론검사(SAT Reasoning Test)의 수준설정에서 패널의 대표성을 유지하기 위하여 최소한 15명 이상으로 하고 모든 패널들이 적극적으로 참여할 수 있도록 30명을 넘지 않도록 권고하기도 한다(Morgan, 2006). 그리고 일반적인 수준설정 절차는 이렇게 구성된 패널들이 장시간 심도 있는 검토와 논의 및 의사결정의 과정을 거쳐 분할점수를 최종적으로 결정하게 된다. 학교현장에서 이러한 정도의 수준설정 절차를 모든 교과에 대해 매 학기마다 적용하는 것은 현실적으로 불가능하다. 또한, 평가도구-중심 수준설정 방법처럼 지필평가와 수행평가의 문항 또는 과제가 모두 정해지거나 시행된 다음에 학교단위에서 소수의 교사들이 분할점수를 결정하도록 하는 것은 많은 부작용을 불러올 수 있다. 학생들의 내신 성적이 상급학교 진학에 중요한 요소가 되는 경우, 교사들은 무의식적으로 학생들의 성취도를 가능한 높게 평정하기 위하여 분할점수를 낮추려는 경향을 보일 가능성이 있을 뿐만 아니라 학부모 및 학교관리자 등 외부의 압력에도 쉽게 노출될 수밖에 없다. 평가도구의 난이도를 낮추어 출제하고 분할점수를 높일 수 있는 모든 권한이 학교의 교사들에게 주어질 경우에 일어날 수 있는 부작용은 성취평가제의 도입 취지를 무색하게 할 수 있다. 또한 수준설정이 지필평가와 수행평가 결과가 이미 나온 이후에 이루어질 경우에는 절대적인 기준에 맞는 점수를 분할점수로 설정하기 보다는 성취도별 적정 학생비율을 대략 정한 다음 이 비율에 준해서 분할점수를 선정할 가능성도 배제할 수 없다.

3. 성취평가제에서 대조집단 수준설정 활용 방안

학교단위에서 지필평가와 수행평가 결과를 토대로 학기말 성취도를 평정하는 절차를 좀 더 체계적으로 관리하는 방안으로 평가도구-중심 수준설정 방법보다는 피험자-중심 수준설정 방법에서 현실적인 대안을 찾는 것이 바람직한 것처럼 보인다. 피험자-중심 수준설정 방법은 지필평

가와 수행평가에 포함되는 구체적인 문항 또는 과제의 내용을 정확하게 알지 않아도 될 뿐만 아니라, 개별 문항 또는 과제의 측정학적 특성을 한 개씩 고려하지 않아도 된다는 측면에서 수준설정 절차를 수행해야 하는 교사의 부담을 크게 줄여줄 수 있다.

현행 성취평가제에서는 한 학기동안 교수·학습과 평가가 모두 이루어진 학생들의 성취도를 평정해야 하는 상황이기 때문에, 피험자-중심 수준설정 방법을 적용한다고 하더라도 현행 성취평가제에서 요구하는 것처럼 학기초에 성취도를 구분하는 분할점수를 미리 결정할 수는 없다. 즉, 교수·학습이 이루어지기도 전에 분할점수를 결정하는 데 필요한 대조집단 또는 경계집단을 단위학교에서 구성하는 것은 불가능할 뿐만 아니라, 설사 그런 집단을 구성할 수 있다고 할지라도 구체적인 평가도구에서 결정되는 분할점수는 최종성적이 산출된 이후에나 정해질 수 있다. 따라서 피험자-중심 수준설정 방법을 사용한다고 하더라도 현행 학업성적관리규정에서 규정하고 있는 것처럼 교과목별 기준성취율(분할점수)를 미리 결정할 수 있는 것은 아니다.

본 연구에서는 학교단위에서 교과별로 한두 명의 교사가 담당교과의 기준성취율의 적정성을 검토해볼 수 있는 방안으로 피험자-중심 수준설정 방법 중에서 대조집단 방법을 제안하고자 한다. 만일 현재의 학업성적관리규정이 개정되어 학기말에 기준성취율을 정하게 된다면 이 방안은 분할점수를 결정하는 원래의 목적으로 사용하면 될 것이다. 피험자-중심 수준설정 방법 중에서 대조집단 방법을 제안하는 것은 성취평가제가 학교단위로 이루어지기 때문에 발생하는 현실적인 문제점들을 고려한 것이다.

피험자-중심 수준설정 방법 중 경계집단 수준설정은 두 성취도 집단의 경계에 있는 학생들을 선택하기 위해 이들의 특징에 대해서 재진술하고 논의하는 것이 필요하다. 그러나 대조집단 수준설정 방법은 학기단위 성취수준기술을 토대로 각 성취도 집단에 속하는 학생들을 선택하면 된다. 대조집단 수준설정 방법은 각 성취수준의 특징을 그대로 사용할 수 있다는 점에서 경계집단 수준설정 방법보다 교사의 부담을 덜어줄 수 있는 장점이 있다. 그리고 교사가 교수·학습 경험을 토대로 성취수준별 학생들을 선택하는 것보다는 성취수준의 경계에 위치하고 있는 학생들을 선택하는 것이 더욱 어려울 뿐만 아니라, 학교단위에서 경계집단의 평균 또는 중앙값의 안정적인 추정치를 얻을 수 있을 만큼 충분한 크기의 표집을 확보하는 것이 어려운 경우도 많이 있다. 또한, 대조집단 수준설정 방법은 평가도구-중심의 수준설정 방법으로 이미 결정된 분할점수가 실제로 타당한지를 검토하는 수단으로도 제안되고 있다(Morgan & Michaelides, 2005). 따라서 성취평가제에서 현행 기준성취율의 타당성을 검토하거나 분할점수를 설정하기 위해 피험자-중심 수준설정 방법 중 한 가지를 선택해야 한다면, 현실적으로 경계집단 방법보다는 대조집단 방법을 사용하는 것이 좀 더 나은 선택일 것이다. 그러나 충분한 수의 학생들이 있는 대규모 학교의 경우에는 경계집단 방법을 사용하는 것도 고려해볼 수 있다.

현행 성취평가제에서 기준성취율의 적정성을 학교단위에서 검토하기 위해서는 평가도구-중심 수준설정 방법을 적용하는 시기가 매우 중요하다. 현실적으로 한두 명의 교사가 교과별 성취도

를 평정하고 이를 검토해야 하는 상황을 고려하면, 교수·학습은 모두 이루어졌으나 학기말 평가는 시행되기 전에 경계집단 또는 대조집단을 선택하는 것이 필요하다. 학교현장에서 학기말 평가가 모두 이루어진 경우에는 교사가 학생들의 학기말 최종 성적을 알게 되므로, 학생들의 수행이 아니라 성취도를 결정하는 학기말 최종 성적을 토대로 경계집단 또는 대조집단을 구성할 가능성이 있다는 측면에서 바람직하지 않다(Zieky & Perie, 2006). 따라서 피험자-중심 수준설정을 위하여 대조집단 또는 경계집단을 선택하는 시점은 한 학기동안의 교수·학습이 거의 완결되고 마지막 학기말 지필평가가 이루어지기 직전이 가장 바람직한 것으로 보인다. 이것은 교사가 한 학기동안 가르치고 평가를 시행하였지만 최종 결과를 모르는 상태이기 때문에 최종 결과를 토대로 인위적으로 분할점수를 조정할 수 있는 가능성을 배제할 수 있는 장치라고 할 수 있다.

Ⅲ. 연구 방법

현행 성취평가제의 기준성취율의 적절성을 평가하는 방안으로 제안한 대조집단 수준설정 방법의 유용성을 검토하기 위하여, 이 방안을 실제 학교현장에 적용해보기 위하여 다음과 같은 연구방법을 사용하였다.

1. 조사대상

○○광역시에 소재하는 중학교 1학년 2개 학급을 선정하고, 해당 학급의 국어, 영어, 수학, 사회, 과학 과목을 가르치는 교사들로 하여금 교수·학습 경험을 토대로 학생들의 성취도를 주관적으로 평가하게 하였다. 국어, 사회, 과학의 경우에는 한 명의 교사가 2개 학급을 모두 가르치고 있었으나, 영어과 수학 과목의 경우에는 3명의 교사가 2개 학급의 학생들을 3개의 집단으로 수준별 수업을 진행하였다. 따라서 영어와 수학 과목의 경우에는 3명의 교사가 실제 수업을 담당하고 있는 학생들의 성취도만을 평정하도록 하였다.

2. 실험 방법 및 절차

성취평가제에서 지필평가와 수행평가의 결과를 종합하여 성취도를 평정하는 기준성취율(분할점수)의 적절성을 평가하기 위하여 다음과 같은 방법과 절차를 사용하여 자료를 수집하였다.

① 조사대상 학교를 선정하고, 자료수집이 용이한 2개 학급을 선정하였다.
② 해당 학급의 교과 담당 교사에게 교과별 학기단위 성취수준의 특성을 검토하고 이를 숙지하도록 하였다.
③ 교사들로 하여금 담당교과의 학생 성취도를 주관적으로 평정하게 하였다. 2개 학급 학생들에 대한 교수·학습 경험을 토대로 학생들의 성취도를 주관적으로 평정할 때 다음 사항을 고려하도록 하였다.
- 교과별 학기단위 성취수준의 특성만을 고려하여 학생들을 평정해야 한다.
- 성취수준의 특성과 관련 없는 요인들이 개입하지 않도록 노력한다.
- 학생들의 수업 이해도, 수업시간 학생활동, 과제물 수행, 쪽지시험, 퀴즈 등에 대한 경험을 토대로 개별 학생의 성취도를 평정한다.
- 학생들의 지필평가 또는 수행평가 성적을 직접적으로 참고하지 않는다.
- 성취도별 학생 비율을 고려하여 평정할 필요는 없으나, 학생들을 서로 비교하면서 성취도를 평정할 수 있다.
④ 지필평가와 수행평가의 점수 반영비율에 따라 100점 만점으로 산출한 학기말 원점수를 토대로 학생들의 학기말 성취도를 평정하였다.
- 영어과 경우에는 1차와 2차 지필평가 성적을 각각 25%씩 반영하고, 수행평가 성적을 50% 반영하여 학기말 원점수를 산출하였다.
- 나머지 교과의 경우에는 1차와 2차 지필평가 성적을 각각 35%씩 반영하고, 수행평가 성적을 30% 반영하여 학기말 원점수를 산출하였다.
⑤ 교사들이 주관적으로 판단한 학기말 성취도와 지필평가와 수행평가의 결과를 토대로 산출된 학기말 성취도의 분류일치도를 분석하였다.
⑥ 성취평가제에서 학기말 성취도를 평정하는 기준이 되는 기준성취율의 적정성을 평가하기 수준설정을 수행하고 기준성취율과 수준설정 분할점수를 비교하였다.

3. 분석방법

교사가 개별 학생에 대해 주관적으로 판단한 성취도 평정 결과와 지필평가와 수행평가의 결과를 종합하여 산정된 학기말 성취도의 일치도를 평가하기 위하여 '분류일치도(classification agreement)'를 추정하였다. 또한, 수준설정 방법 중에 피험자를 기준으로 분할점수를 결정하는 대조집단 방법을 사용하여 학기단위 성취수준의 분할점수를 설정하고, 이를 토대로 실제로 학기단위 성취수준을 구분하기 위해 사용한 기준성취율의 적절성을 평가하였다.

분류일치도는 동일한 대상들을 2명의 채점자, 2회의 검사, 또는 다른 2가지 방법을 사용하여

독립적으로 분류하였을 때, 그 분류가 얼마나 일치하는지를 나타내는 척도라고 할 수 있다. 식 (1)은 두 가지 분류의 일치도를 나타내는 코헨의 카파계수(Cohen's kappa coefficient)를 추정하는 공식이다.

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e} \quad \text{단, } P_0 = \sum_{i=1}^{5} p_{ii}\,,\ P_e = \sum_{i=1}^{5} p_{.i} p_{i.} \qquad \text{식 (1)}$$

식 (1)에서 분자는 두 가지 분류가 일치하는 확률(P_o)이 우연에 의해 일치하는 확률(P_e)을 제외하고 얼마나 큰 지를 나타내고, 분모의 경우에는 확률 1에서 우연에 의해 일치하는 확률을 제외한 크기를 나타낸다. 따라서 κ 는 우연을 넘어서서 일치하는 전체 확률에서 우연을 넘어서서 일치하는 관찰된 확률의 비를 나타낸다(Cohen, 1960). 두 가지 방법으로 분류한 것이 모두 정확하게 일치하는 경우 $\kappa = 1$이 된다. 두 가지 방법으로 분류한 것이 일치할 확률이 우연에 의해 일치하는 확률보다 클 경우에는 양수가 되고 그렇지 않을 경우 음수가 되지만, κ 의 값은 양수가 되는 것이 일반적이다. 한편, 식 (2)는 일치하지 않는 정도에 가중치를 두어 분류일치도를 정의한 가중 카파계수 (weighted kappa coefficient, $\kappa_{(w)}$)를 보여주고 있다(Cohen, 1968). 이것은 두 가지 분류의 일치여부만 따지는 것이 아니라, 일치하지 않는 정도를 정의하고 이를 분류일치도 계수에 반영한 것이라고 할 수 있다.

$$\hat{\kappa}_{(w)} = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}} \quad \text{단, } P_{o(w)} = \sum_{i=1}^{5}\sum_{j=1}^{5} w_{ij} p_{ij}\,,\ P_{e(w)} = \sum_{i=1}^{5} w_{ij} p_{.i} p_{i.} \qquad \text{식 (2)}$$

식 (3)은 피어슨 χ^2 검정 통계량을 계산하는 공식이다. 이 공식은 두 변인의 독립성을 가정하고 계산한 기댓값과 관찰값 간 차이의 유의미성을 검정하기 위한 통계량이다. 두 변인이 관련성이 없는 경우에는 기댓값과 관찰값 사이에 차이가 작아 Q_P의 값이 작아지겠지만, 어떤 형태로든 관련성이 높을수록 그 차이가 커져서 Q_P의 값은 작아지게 된다. 그러나 Q_P의 범위가 정해질 수 있는 것이 아니기 때문에 Q_P의 절대적인 크기로 두 분류의 관련성을 해석하는데 한계가 있다.

$$Q_P = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad \text{단, } e_{ij} = \frac{n_{i.} n_{.j}}{n} \qquad \text{식 (3)}$$

식 (4)는 이 Q_P의 값을 보정하여 표준화시킨 크레머 V 계수(Cramer's V coefficient)를 계산하는 공식이다. 크레머 V 계수는 항상 0보다 크거나 같고 1보다 작거나 같은 값을 가지기 때문에 두 분류의 관련 정도를 해석하는데 Q_P보다 유용하다.

$$V=\sqrt{\frac{Q_p/n}{\min(R-1,C-1)}} \quad \text{단, } 0 \leqq V \leqq 1 \qquad \text{식 (4)}$$

Ⅳ. 분석 결과 및 논의

1. 분류일치도

교사가 주관적으로 평정한 학생들의 성취도와 지필평가 및 수행평가를 토대로 평정한 성취도가 어느 정도 일치하는지를 분석하였다. 이를 위하여 분류교차표를 제시하였고, 분류일치도와 분류 상관계수를 추정하였다. 분류교차표는 동일한 학생의 두 가지 성취도를 교차시켜 만든 표로서 두 가지 성취도가 얼마나 일치하는지를 확인할 수 있다. 이 분류교차표에서 단순하게 두 개의 성취도가 정확하게 일치하는 비율로 분류일치도를 추정하였다. 그러나 이 분류일치도는 이해가 매우 쉽지만 우연에 의해 두 개의 성취도가 일치하는 정도를 배제하지 못하는 문제점이 있다. 따라서 우연에 의해 두 개의 성취도가 일치하는 정도를 배제한 분류일치도 계수를 별도로 추정하였다. 또한, 범주형 변인에 해당하는 두 개의 성취도 간에 분류 상관관계를 추정하였다.

가. 분류교차표와 분류일치도

〈표 1〉~〈표 5〉는 교과별로 교사가 주관적으로 평정한 성취도와 평가결과를 토대로 평정한 성취도가 어느 정도 일치하는지를 보여주는 분류교차표이다. 〈표 1〉은 국어과에서 교사가 주관적으로 평정한 성취도와 평가결과를 토대로 평정한 성취도를 비교한 결과 77명의 학생 중에서 30명의 성취도 평정 결과가 정확하게 일치하여, 두 가지 성취도의 분류일치도는 39%인 것으로 나타났다. 〈표 2〉는 영어과의 성취도 평정 결과의 일치도를 보여주고 있는데, 77명의 학생 중 50명의 성취도 평정 결과가 정확하게 일치하는 것으로 나타났다. 영어과 성취도의 분류일치도는 65%인 것으로 나타났다.

〈표 3〉은 수학과에서 교사가 주관적으로 평정한 성취도와 평가결과를 토대로 평정한 성취도를 비교한 결과, 77명의 학생 중 48%에 해당하는 37명의 성취도 평정 결과가 정확하게 일치하는 것으로 나타났다. 〈표 4〉는 과학과의 성취도 평정 결과의 일치도를 보여주고 있는데, 77명의 학생 중 82%에 해당하는 63명의 성취도 평정 결과가 정확하게 일치하는 것으로 나타났다.

〈표 1〉 국어과 성취도 교차표

국어		교사가 판단한 성취도					계
		A	B	C	D	E	
실제성취도	A	9	0	0	0	0	9
	B	11	8	0	0	0	19
	C	1	11	2	0	0	14
	D	0	4	12	1	0	17
	E	0	0	4	4	10	18
계		21	23	18	5	10	77

〈표 2〉 영어과 성취도 교차표

영어		교사가 판단한 성취도					계
		A	B	C	D	E	
실제성취도	A	25	5	0	0	0	30
	B	7	6	0	0	0	13
	C	2	5	3	1	0	11
	D	0	2	3	6	0	11
	E	0	1	0	1	10	12
계		34	19	6	8	10	77

〈표 3〉 수학과 성취도 교차표

수학		교사가 판단한 성취도					계
		A	B	C	D	E	
실제성취도	A	16	0	0	0	0	16
	B	8	5	0	1	0	14
	C	3	9	2	0	0	14
	D	0	2	4	4	0	10
	E	2	1	5	5	10	23
계		29	17	11	10	10	77

〈표 4〉 과학과 성취도 교차표

과학		교사가 판단한 성취도					계
		A	B	C	D	E	
실제성취도	A	16	2	0	0	0	18
	B	0	16	2	0	0	18
	C	0	2	10	2	0	14
	D	0	1	1	7	0	9
	E	0	0	0	4	14	18
계		16	21	13	13	14	77

〈표 5〉는 사회과에서 교사가 주관적으로 평정한 성취도와 평가결과를 토대로 평정한 성취도를 비교한 결과 77명의 학생 중에서 36%에 해당하는 28명의 성취도 평정 결과가 정확하게 일치하는 것으로 나타났다.

〈표 5〉 사회과 성취도 교차표

사회		교사가 판단한 성취도					계
		A	B	C	D	E	
실제성취도	A	6	8	0	0	0	14
	B	1	10	2	0	0	13
	C	0	6	4	0	0	10
	D	0	2	9	2	0	13
	E	0	0	8	13	6	27
계		7	26	23	15	6	77

〈표 6〉은 교사가 주관적으로 판단한 성취도와 평가결과를 토대로 평정한 성취도가 우연에 의해 일치하는 정도를 제외한 분류일치도 κ와 $\kappa_{(w)}$의 추정치와 각각의 95% 신뢰구간을 보여주고 있다. 분류일치도가 가장 높은 과목은 과학인 것으로 나타났고, 이어서 영어, 수학, 국어, 사회 과목의 순으로 나타났다.

과학과의 분류일치도 $\kappa = 0.77$은 우연에 의해 두 개의 성취도가 일치할 수 있는 부분을 제외한 나머지 부분의 77%가 추가로 일치한다는 것을 의미한다. 이 분류일치도에 대한 95% 신뢰구간은 0.66보다 크고 0.88보다 작은 것으로 나타났다. 불일치하는 정도에 따라 가중치를 부여한 분류일치도는 $\kappa_{(w)} = 0.88$인 것으로 나타났는데, 불일치 정도에 가중치를 부여하여 확률을 계산하는 경우에 우연에 의해 일치하는 부분을 제외한 나머지 부분의 88%가 추가로 일치한다는 것을 의미한다. 이 분류일치도에 대한 95% 신뢰구간은 0.82보다 크고 0.94보다 작은 것으로 나타났다.

한편, 분류일치도가 가장 낮은 사회과의 경우, $\kappa = 0.23$은 우연에 의해 두 개의 성취도가 일치할 수 있는 부분을 제외한 나머지 부분의 23%가 추가로 일치한다는 것을 의미한다. 이 분류일치도에 대한 95% 신뢰구간은 0.11보다 크고 0.35보다 작은 것으로 나타났다. 불일치 정도에 따라 가중치를 부여한 분류일치도 $\kappa_{(w)} = 0.52$는 우연에 의해 일치하는 부분을 제외한 나머지 부분의 52%가 추가로 일치한다는 것을 의미한다. 이 분류일치도에 대한 95% 신뢰구간은 0.43보다 크고 0.62보다 작은 것으로 나타났다.

영어, 수학, 국어과의 κ 추정치는 우연에 의해 두 개의 성취도가 일치할 수 있는 부분을 제외한 나머지 부분의 53%, 35%, 24%가 각각 더 일치하는 것으로 나타났고, 불일치 정도에 따라 가중치를 부여한 $\kappa_{(w)}$ 추정치는 우연에 의해 두 개의 성취도가 일치할 수 있는 부분을 제외한 나머지 부분의 73%, 57%, 55%가 각각 더 일치하는 것으로 나타났다.

〈표 6〉 교과별 성취도 분류일치도

분류일치도 계수		추정치	표준오차	95% 신뢰구간	
				하한치	상한치
κ	국어	0.24	0.07	0.11	0.37
	영어	0.53	0.07	0.39	0.67
	수학	0.35	0.07	0.22	0.48
	과학	0.77	0.05	0.66	0.88
	사회	0.23	0.06	0.11	0.35
$\kappa_{(w)}$	국어	0.55	0.05	0.45	0.65
	영어	0.73	0.05	0.63	0.83
	수학	0.57	0.06	0.45	0.68
	과학	0.88	0.03	0.82	0.94
	사회	0.52	0.05	0.43	0.62

교과별 분류일치도 κ가 어느 정도의 일치 수준을 나타내는지를 비교해보면, 과학과 영어 교과의 분류일치도 수준은 각각 실질적인 수준의 일치(substantial agreement)와 보통수준의 일치(moderate agreement)라고 할 수 있으며, 나머지 교과의 경우에는 상당한 수준의 일치(fair agreement)라고 할 수 있다(Landis & Koch, 1977).

나. 분류 상관관계

〈표 7〉은 교사의 주관적 판단에 따라 평정한 성취도와 평가결과를 토대로 평정한 성취도와의 상관관계에 해당하는 크레머 V 계수 추정치를 보여주고 있다. 피어슨 χ^2 검정 통계량은 두 가지 방법으로 평정한 성취도 사이에 아무런 관련이 없을 때 두 가지 성취도 조합 각각에 나타나는 경우의 수에 비하여 관찰된 경우의 수가 얼마나 큰지를 나타낸다. 크레머 V 계수는 피어슨 χ^2 검정 통계량를 보정하여 최솟값이 0이 되고 최댓값이 항상 1이 되도록 표준화시킨 것이다.

과학과의 상관관계가 가장 높게 나타났고 사회과의 상관관계가 가장 낮은 것으로 나타났다. 과학과의 경우 $V=0.78$인 것으로 나타났는데, 교사가 주관적으로 평정한 성취도와 평가결과를 토대로 평정한 성취도 간 아무런 관련이 없다고 가정하였을 때 기대되는 교차표의 빈도수와 실제로 관찰된 교차표의 빈도수 간에 차이의 정도를 수치로 표현하였을 때 가장 높은 값의 약 78%에 해당하는 값을 갖는다는 의미이다. 한편, 사회과의 분류 상관계수 추정치는 $V=0.53$인 것으로 나타났는데, 이것은 가장 높은 분류 상관계수 수준의 약 53%에 해당한다는 것을 의미한다. 수학과의 분류 상관계수는 사회과와 유사한 수준인 것으로 나타났다. 한편, 국어과와 영어과의 분류 상관계수 추정치는 서로 매우 유사하였는데, 가장 높은 값의 약 63%에 해당하는 것으로 나타났다.

〈표 7〉 교과별 크레머 V 계수

교과	국어	영어	수학	과학	사회
V	0.63	0.63	0.55	0.78	0.53

2. 대조집단(contrast groups) 방법을 사용한 수준설정

〈표 8〉~〈표 12〉는 교사가 주관적으로 평정한 성취도를 기준으로 구분한 대조집단들의 학기말 원점수 분포의 중앙값, 평균, 표준편차를 보여주고 있다. 교사가 2개 학급 76명의 학생들을 5개의 성취도 집단으로 범주화할 때, 개별 학생의 수행 수준을 고려하였을 뿐 집단별 비율을 고려하지는 않았다. 따라서 대조집단의 크기가 균등하게 나타나지는 않았는데, 모든 과목에서 학생 수가 10명 이하인 대조집단이 나타났다. 대조집단의 크기가 충분히 크지 않아 극단적으로

높거나 낮은 점수가 집단의 평균에 미치는 영향이 매우 클 경우에는 집단을 대표하는 값으로 평균값보다는 중앙값을 사용하는 것이 바람직하다(Livingston & Zieky, 1982). 〈표 8〉~〈표 12〉에 극단적인 점수의 영향으로 중앙값과 평균 사이에 큰 차이가 있는 경우들이 있다. 예를 들어, 사회과 A 집단은 크기는 7명인데 평균값이 중앙값보다 10.9점 낮은 것으로 나타났다.

〈표 8〉 국어과 성취도 집단별 평균과 표준편차

구분 \ 성취도	A	B	C	D	E	전체
빈도 수	21	22	18	5	10	76
중앙값	86.2	77.0	65.8	53.7	45.6	74.7
평균	80.5	78.0	68.3	55.9	47.0	70.9
표준편차	16.3	9.3	10.7	6.8	19.6	17.4

〈표 9〉 영어과 성취도 집단별 평균과 표준편차

구분 \ 성취도	A	B	C	D	E	전체
빈도 수	34	18	6	8	10	76
중앙값	94.5	80.6	73.1	68.4	34.7	84.9
평균	89.1	78.5	72.8	72.1	44.0	77.6
표준편차	14.0	15.0	8.3	11.4	29.7	21.8

〈표 10〉 수학과 성취도 집단별 평균과 표준편차

구분 \ 성취도	A	B	C	D	E	전체
빈도 수	28	17	11	10	10	76
중앙값	89.0	73.3	72.9	56.3	25.4	74.1
평균	81.2	72.7	68.9	57.1	36.8	68.5
표준편차	19.8	14.3	12.9	13.0	24.8	22.7

〈표 11〉 과학과 성취도 집단별 평균과 표준편차

구분 \ 성취도	A	B	C	D	E	전체
빈도 수	15	21	13	13	14	76
중앙값	93.2	83.2	77.8	65.3	43.9	76.9
평균	88.3	78.9	79.2	67.1	46.4	72.8
표준편차	14.9	16.6	8.9	10.3	19.2	20.2

〈표 12〉 사회과 성취도 집단별 평균과 표준편차

구분 \ 성취도	A	B	C	D	E	전체
빈도 수	7	25	23	15	6	76
중앙값	92.9	82.4	64.5	48.1	25.2	69.1
평균	82.0	80.0	63.1	54.8	33.0	66.4
표준편차	29.4	13.6	12.7	18.0	25.8	22.0

대조집단을 이용한 수준설정 방법은 대조집단의 점수 분포를 고려하여, 전체적으로 오분류 확률이 가장 낮은 점수를 분할점수로 결정한다. 구체적으로, 분할점수를 결정하는 방법으로는 두 집단의 관찰점수의 분포 곡선을 그리고 두 곡선이 교차하는 지점을 찾는 방법, 간단하게 두 집단의 중앙값 또는 평균의 중간값을 선택하는 방법, 로지스틱 회귀분석 방법을 적용하여 확률적으로 분할점수를 선택하는 방법이 있다(Cizek & Bunch, 2007). 그러나 이 연구에서는 대조집단별 표본의 수가 충분히 크지 않기 때문에 각 성취도 집단별 점수분포의 대푯값으로 중앙값을 선택하여 분할점수를 결정하였다. 즉, 교과별로 이웃하는 성취도 집단의 중앙값의 중간값을 두 성취도 집단을 구분하는 분할점수로 결정하였다. 〈표 13〉은 대조집단의 중앙값을 이용한 교과별 분할점수와 기준성취율[2]과 분할점수의 차이를 보여주고 있다. 구체적으로 대조집단을 이용한 수준설정에서는 국어와 수학의 경우에는 A와 B를 구분하는 분할점수가 약 81점인 것으로 나타났으나, 나머지 과목의 경우에는 87~88점인 것으로 나타났다. 한편, 중학교 학생들의 학기말 성취도 평정에서는 기준성취율을 적용하여 90점 이상인 학생들의 성취도만 A로 평정하였다. 이것은 A와 B를 분할하는 기준성취율이 성취도를 구분하기에 2~9점 정도 높은 수준이라는 것을 의미한다. 〈표 13〉에서 대조집단을 이용한 분할점수가 기준성취율보다 낮은 것으로 나타난 것은 영어과의 C와 D를 구분하는 분할점수와 과학과의 C와 D 및 D와 E를 구분하는 분할점수인 것으로 나타났고, 다른 경우에는 모두 기준성취율이 분할점수보다 높은 것으로 나타났다.

〈표 13〉 교과별 분할점수와 기준 성취율과의 차이

교과	분할점수				기준성취율−분할점수				
	A/B	B/C	C/D	D/E	A/B	B/C	C/D	D/E	평균
국어	81.6	71.4	59.7	49.7	8.4	8.6	10.3	10.4	9.4
영어	87.5	76.9	70.8	51.5	2.5	3.2	-0.8	8.5	3.4
수학	81.1	73.1	64.6	40.8	8.9	7.0	5.5	19.2	10.1
과학	88.2	80.5	71.5	54.6	1.8	-0.5	-1.5	5.4	1.3
사회	87.6	73.4	56.3	36.7	2.4	6.6	13.7	23.3	11.5

2) 중학교 성취평가제에서 학기말 성취도 A, B, C, D, E를 구분하는 평정하는 기준은 각각 90, 80, 70, 60점이다.

3. 요약 및 논의

교사가 교수·학습 경험을 토대로 주관적으로 평정한 성취도와 실제 평가결과를 토대로 평정한 성취도가 일치하는 정도는 교과에 따라 차이가 있었다. 우선, 두 가지 성취도가 정확하게 일치하는 비율을 의미하는 단순 분류일치도는 36%~82%인 것으로 나타났다. 그리고 코헨의 분류일치도 κ는 우연에 의해 두 성취도가 정확하게 일치하는 것을 제외한 나머지 부분의 23%~77%가 더 정확하게 일치하는 것으로 나타났다. 또한, 일치하지 않는 정도에 따라 가중치를 부여하여 계산한 코헨의 분류일치도 $\kappa_{(w)}$는 우연에 의해 일치하는 것을 제외한 나머지 부분의 52%~88%가 더 일치하는 것으로 나타났다. 두 가지 성취도가 정확하게 일치하는 정도를 따졌을 때는 과목별로 큰 차이가 날 뿐만 아니라 과학과 영어를 제외한 과목에서 50% 이하의 일치도를 보였다. 그러나 불일치하는 정도에 가중치를 부여하여 계산한 분류일치도의 경우에는 모든 과목이 50% 이상의 일치도를 보였다. 이것은 두 가지 성취도가 불일치하는 경우에도 등급의 차이가 그렇게 크지 않다는 것을 의미한다. 그리고 분류교차표는 최상위 및 최하위 성취수준에 있는 학생들의 경우 두 성취도가 상당히 일치하는 것을 보여주고 있으며, 중간 성취수준에 있는 학생들은 상대적으로 그 차이가 다소 큰 것으로 나타났다.

교과별 분류교차표를 살펴보면, 3개 교과(국어, 수학, 사회)의 성취도 오분류 양상이 비대칭인 것으로 나타났다. 교사가 주관적으로 평정한 성취도를 기준으로 보았을 때, 평가결과를 토대로 평정한 성취도의 편차가 한 쪽으로 기우는 경향이 있다는 것을 의미한다. 교과 및 성취도 구간에 따라 다소의 차이가 있지만, 교사가 주관적으로 평정한 성취도에 비하여 평가결과를 토대로 평정한 성취도가 전반적으로 낮은 경향이 있었다. 이것은 교사가 의도했던 것보다 지필평가와 수행평가의 난이도가 특정 성취수준의 학생들에게 어려울 수 있었다는 것을 의미한다. 이와는 반대로 지필평가와 수행평가의 난이도가 적절했다면, 교사들의 학생들의 성취도를 다소 높게 평정하는 경향이 있다는 해석도 가능하다. 분류교차표와 분류일치도에 대한 분석 결과에서 얻을 수 있었던 것은 성취도를 구분하는 분할점수를 미리 고정시키더라도 교사의 전문성과 경험을 토대로 어느 정도 적절한 평가가 이루어질 수 있다는 것을 의미한다. 그러나 이것은 교과에 따라 큰 차이가 있었는데, 교사 및 교과의 특성뿐만 아니라 교과별 평가방법의 차이 등이 원인이 될 수 있다. 또한, 분류일치도가 높은 교과의 경우에도 성취도 구간에 따라 일치하는 정도에 편차가 다르게 나타났다. 이것은 성취도를 구분하는 분할점수가 모든 성취도 구간에서 적절한 것은 아니라는 것을 의미한다.

성취평가제에서 성취도를 구분하는 분할점수인 기준성취율의 적절성을 평가하기 위하여 교사가 주관적으로 평정한 성취도를 기준으로 대조집단을 구성하여 수준설정을 하였다. 이를 위하여, 학기초에 미리 고정시켜 놓은 기준성취율이 학기말 평가가 이루어진 후에 대조집단 수준설정 방법을 사용하여 결정한 분할점수와 어느 정도 일치하는지를 기준성취율 평가의 준거로 삼았

다. 교과별로 기준성취율과 분할점수의 차이를 보면, 모든 교과에서 기준성취율이 분할점수보다 평균적으로 높은 것으로 나타났다. 과학과의 경우에는 기준성취율과 분할점수의 차이가 평균적으로 1.3점 정도밖에 나지 않았으며, D와 E 수준을 구분하는 분할점수를 제외하고는 그 차이가 2점 이하인 것으로 나타났다. 영어과의 분할점수는 과학과의 분할점수와 매우 유사한 양상으로 나타났는데, 그 차이가 조금 더 크다는 차이점이 있었다. 한편, 국어과의 경우에는 모든 성취도 구간에서 기준성취율과 분할점수의 차이가 8~10점 정도인 것으로 나타났는데, 이것은 모든 성취도 수준에 있는 학생들에게 지필평가와 수행평가가 교사가 의도했던 것보다 일정하게 더 어려웠다는 것을 의미한다. 그리고 수학과와 사회과의 경우에는 전반적으로 분할점수와 기준성취율의 차이가 상대적으로 크게 나타났다. 모든 성취도 구간에서 기준성취율이 분할점수보다 높게 나타났는데, 특히 D와 E를 구분하는 분할점수의 차이가 20점 내외인 것으로 나타났다. 이것은 D와 E 집단을 구분하는 기준성취율이 적정수준보다 20점 정도 높다는 것을 의미하는데, 교사가 의도했던 것보다 낮은 성취도 수준의 학생들에게 매우 어려운 수준의 평가이었음을 추정할 수 있었다. 이와 같은 대조집단 수준설정 방법은 이미 평가도구-중심 수준설정 방법으로 결정한 분할점수의 타당성을 검토하는 방안으로 제안되고 있으며(Morgan & Michaelides, 2005), 현행 성취평가제의 분할점수에 해당하는 기준성취율의 타당성을 검토함으로써 한 학기 동한 시행된 지필평가와 수행평가의 난이도 수준이 적절했는지를 평가할 수 있는 수단으로도 사용될 수 있다. 한편, 이 대조집단 수준설정 방법에서는 패널들이 피험자들의 수행수준을 얼마나 잘 판단해서 대조집단을 구성하는지가 매우 중요하다(Morgan & Michaelides, 2005). 대조집단 수준설정 방법에서도 다수의 패널들이 대조집단을 구성하는 것이 일반적이지만, 이 연구에서는 한 학기동한 학생들을 직접 가르친 교사가 자신의 교수·학습 경험을 토대로 학생들을 구분하여 대조집단을 구성하였다. 성취평가제는 학교단위에서 이루어지는 것으로 한 학기동안 학생들을 직접 가르쳐온 교사만큼 대조집단을 잘 구성할 수 있는 패널들을 구성하는 것은 어려워 보인다. 즉, 다수의 패널이 대조집단을 구성한 것은 아니지만, 한 학기 동안 학생들을 직접 가르치고 관찰해온 교사가 대조집단을 구성하였다는 점에서 무리가 없어 보인다. 이와 같은 관점에서 이 연구에서 제안하고 있는 대조집단 수준설정 방법은 학교현장에서 성취평가제의 기준성취율의 타당성을 검토함으로써, 교사의 업무 부담을 크게 늘이지 않으면서 한 학기동안 수행한 평가의 난이도 수준을 모니터링할 수 있는 수단으로 사용할 수 있다는 점에서 현실적으로 적절하고 유용한 방안이라고 할 수 있다.

또한, 향후 학기말에 수준설정 방법을 사용하여 분할점수를 결정할 수 있도록 허용될 경우에도, 학교 현장에서 매 학기 모든 교과에서 대표성 있는 20~30명의 패널을 구성해서 장시간의 토론과 의사결정 과정을 거쳐야 하는 평가도구-중심 수준설정 절차를 수행하기는 매우 어렵다. 그런데 한 학기동안 학생들에 대한 교수·학습 경험을 토대로 대조집단을 구성해서 분할점수를 결정하는 방안은 학생들의 수행 특성을 잘 아는 교사가 대조집단을 구성한다는 점에서 적절하

고, 교사의 업무 부담을 크게 늘이지 않고 분할점수를 결정할 수 있다는 점에서 현실적으로 유용하며, 분할점수 결정과 관련된 외부의 압력의 상대적으로 많이 피할 수 있다는 점에서 현실성 있는 수준설정 방안이 될 것이다.

Ⅴ. 결론 및 제언

이 연구는 현행 성취평가제에서 성취도 평정의 기준으로 사용하고 있는 기준성취율의 적정성을 검토하기 위하여 학교단위에서 사용할 수 있는 방안으로 피험자-중심의 수준설정 방법을 제안하였다. 특히, 피험자-중심의 수준설정 방법 중에서도 대조집단 방법이 학교단위에서 피험자 집단을 구성하기가 용이하고 교사의 업무부담을 줄일 수 있을 뿐만 아니라, 이미 다른 수준설정 방법을 통해 결정한 분할점수의 타당도를 점검하는 수단으로 사용되고 있다는 점에서 우선적으로 고려할 필요가 있음을 강조하였다. 그리고 학교단위에서 대조집단 방법을 사용하여 기준성취율의 적정성을 평가하는 방안의 유용성을 보여주기 위하여 실제 학교현장에 실험적으로 적용하여 보았다.

우선, 교사가 교수·학습 과정에서 학생들을 관찰한 경험을 토대로 2개 학급 학생들의 성취도를 주관적으로 평정하였다. 교사가 주관적으로 평정한 성취도는 지필평가와 수행평가를 토대로 한 학기말 성취도와 비교되었으며, 분할점수 결정을 위해 대조집단을 구성하는 기준이 되었다. 교사가 주관적으로 평정한 성취도 결과와 평가결과를 토대로 평정한 성취도 결과가 일치하는 정도는 교과에 따라 다소 큰 차이가 있었다. 또한 대체적으로 가장 높은 성취수준 A와 가장 낮은 성취수준 E는 일치하는 경향이 있었으나, 성취수준 B의 경우에는 불일치하는 정도의 폭이 다소 넓은 경향이 있었다. 대조집단 수준설정 방법으로 결정한 성취수준 분할점수와 기준성취율의 차이도 교과와 성취수준에 따라 다소 큰 폭의 차이가 있었다. 국어와 수학은 성취수준 전체에 걸쳐서 전반적으로 큰 폭의 차이가 나타났고, 사회는 하위 성취수준의 분할점수에서 큰 차이가 나타났으며, 모든 교과에서 D와 E 성취수준을 구분하는 분할점수에서 큰 차이가 나타났다. 이것은 교과에 따라 차이가 나타났지만, 교사들이 성취도를 구분하는 분할점수가 사전에 정해진 경우에도 학생들의 성취도가 적절하게 평정될 수 있도록 지필평가와 수행평가의 난이도를 어느 정도 조절할 수 있다는 것을 의미한다.

그럼에도 불구하고, 현행 성취평가제처럼 학기말 성취수준을 구분하는 분할점수를 미리 고정시켜 놓고, 교사의 경험과 전문성을 토대로 지필평가와 수행평가의 난이도를 조절하도록 하는 것은 학기말 성취도 평정에 대한 교사의 심리적 부담을 크게 가중시키고 있다. 교사의 경험과 전문성만으로 5개의 성취수준을 적절하게 구분할 수 있도록 지필평가와 수행평가의 문항 또는

과제의 난이도를 조정하도록 하는 것은 통제할 수 있는 한계를 벗어나는 것이라고 할 수 있다. 교사가 학생들의 성취도를 지나치게 나쁘게 평가하는 실수를 피하기 위해서 지필평가와 수행평가를 최대한 쉽게 출제하는 결과를 초래할 수 있다. 따라서 학생들의 학기말 성취도를 평정할 수 있는 합리적인 절차를 제공함으로써 학생들의 학기말 성취도 평정이 적절하게 이루어질 수 있도록 하고, 미리 정해진 성취도 분할점수에 맞추어 지필평가와 수행평가의 난이도를 조정해야 하는 부담을 덜어주는 것이 필요하다. 이를 위해서는 학생들의 성취수준을 구분하는 분할점수를 학기 초에 결정하도록 되어 있는 학업성적관리규정이 먼저 개정되어야 한다. 그러나 지필평가와 수행평가가 모두 이루어진 이후에 교수·학습과 평가를 직접 담당한 소수의 교사들이 평가도구-중심 수준설정 방법을 사용하는 경우, 분할점수 결정 시 성적 부풀리기와 같은 부작용을 야기할 수 있다. 따라서 현행 학업성적관리규정처럼 기준성취율을 사전에 미리 규정하고 사후에 피험자-중심 수준설정 방법을 적용하여 어느 정도까지 분할점수를 보정할 수 있는 여지를 만들어 주는 것이 필요하다.

이 연구에서 성취평가제의 기준성취율의 적절성을 평가하는 방안으로 피험자-중심의 대조집단 수준설정 절차를 적용하는 방안을 제안하였으나, 향후에 평가도구-중심의 수준설정 방법과의 결과를 비교하거나 국가수준 학업성취도 평가와 같은 표준화 검사와의 결과를 비교하는 방안에 대한 후속연구가 필요하다.

참 고 문 헌

교육과학기술부(2011). 창의・인성교육 강화를 위한 「중등학교 학사관리 선진화 방안」 발표: 고교 석차 9등급제 평가를 성취평가제로 전환. 보도자료(2011.12.14.).

한국교육과정평가원(2012a). **2012학년도 성취평가제 운영 매뉴얼: 중학교용**. 한국교육과정평가원 연구자료 ORM 2012-18.

한국교육과정평가원(2012b). **2012학년도 성취평가제 운영 매뉴얼: 특성화고・마이스터고・종합고 전문교과용**. 한국교육과정평가원 연구자료 ORM 2012-19.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development,* 225-258. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting*. Thousands Oak, CA: SAGE.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed.). Westport, CT: American Council on Education/Praeger.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.), 485-514. New York: Macmillan.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Morgan, D. L. (2006). *Setting local cut scores on the SAT Reasoning Test Writing Section* (College Board Special Report). New York: The College Board.

Morgan, D. L., & Michaelides, M. P. (2005). *Setting Cut Scores for College Placement* (College Board Research Report No. 2005-9). New York: The College Board.

Thorndike, R. M. (1997). *Measurement and Evaluation in Psychology and Education*. New Jersey: Prentice-Hall.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.

· *논문접수 : 2012-09-01/ 수정본접수 : 2012-10-11/ 게재승인 : 2012-10-17*

ABSTRACT

How to evaluate the appropriateness of cut scores to determine students' achievement levels based on the criterion-referenced evaluation

Sang-Ha Lee
(Research Fellow, Korea Institute for Curriculum and Evaluation)
Hyuk-Joon Choi
(Research Fellow, Korea Institute for Curriculum and Evaluation)

The purpose of this study is to suggest how the secondary school teachers can review if the difficulty levels of their tests were appropriate to evaluate students' achievement levels according to the predetermined cut scores. Ministry of Education, Science, and Technology introduced the criterion-referenced evaluation system into secondary schools in 2012.

To verify utility of the suggested method, we selected two classes of 7th grade and asked teachers to judge the students' achievement levels based on their teaching experience without referring to students' any test scores at the end of semester. The subjective judgement of students' achievement levels were compared with their final achievement levels based on their test scores. The accuracy of classification between the both achievement levels were estimated. The kappa coefficient varied from 0.23 to 0.77, and the weighted kappa coefficient did from 0.52 to 0.88. In addition to this analysis, the cut scores predetermined at the beginning of semester were compared with the cut scores determined by contrast group method. The predetermined cut scores were higher than the cut scores of the standard setting by 1.3~11.5 points. Teachers were quite good at adjusting the difficulty levels of their paper-pencil test and performance assessment for the predetermined cutscores. Using the contrast group method was suggested to review appropriateness of the cut scores determined at the beginning of semester or to determine the cut scores at the end of semester.

Key Words : standard setting, contrast groups method, cut scores, criterion-referenced evaluation