교육과정평가연구
The Journal of Curriculum and Evaluation
2012, Vol. 15, No. 1, pp. 25~50

교과서 검정 심사의 분류일치도 분석 연구[1)]

김 창 환(한국교육과정평가원 성과평가실장)*

≪ 요 약 ≫

현재 우리나라에서는 피험자들이 검사에서 통과하거나 실패하는 여부는 어떠한 과학적이고 체계적인 기준 설정의 과정이 없이 임의로 정한 점수 평균을 넘어서는가를 재는 동일한 기준에 따라 보통 결정되는데, 최근 점수 판정제가 도입된 '교과서 검정 심사'에서도 이러한 분할점수 설정이 적용되고 있으며 분류 결과에 대한 분류일치도 분석 및 추정은 아직까지 시도된 바가 없었다. 이러한 상황에서, 이 연구는 복합 문항(교과서 검정 심사 결과)에 대하여 세 가지의 분류일치도 추정 방식을 사용하여 도출된 분류일치도 분석을 목표로 하였으며 자료는 2006년 개정 교육과정에 의한 '2008년도 수학, 영어 과목의 교과서 검정 심사' 결과 자료를 사용하였다. 연구 결과를 종합한 결론은 첫째, 분류일치도를 추정하는 다섯 가지의 방식 중 세 가지 즉, NM 방식, LL 방식, CM 방식을 사용하여 분류일치도를 추정하였는데 세 가지 방식에서 산출된 P 계수와 Kappa 계수의 추정치들은 모두 다소 높은 수준을 보였다. 둘째, 모형을 적용한 방식의 가정 적합성 확인을 위하여 NM 방식의 정상성 가정은 Q-Q 플롯을 그려본 결과 모든 점들이 거의 직선을 따라 놓여있는 것을 확인하였으며, LL 방식의 4모수 베타-이항 모형은 불규칙성이 많이 보였고 전체적으로 관찰점수 및 조정점수의 분포는 대략 동일한 분포 형태를 갖는 것으로 보였다. 마지막으로, 교과서 검정에서 채택될 수 있는 여러 가지 제언들이 요약되어 제안되었고 연구의 한계 및 장래의 연구 가능성을 제시하였다.

주제어 : 준거참조검사, 분류일치도, P 계수, Kappa 계수, NM 방식, LL 방식, CM 방식, 교과서 검정 심사

1) 이 논문은 저자의 박사학위 취득 논문(고려대학교 대학원 교육학과, 2009)에서 일부를 발췌・수정하여 작성되었음.

* 제1저자 및 교신저자, kimch@kice.re.kr

Ⅰ. 서론

학생의 성취 결과를 해석하고 보고하는 방식은 일반적으로 크게 두 가지로 구분할 수 있는데, 준거참조평가(criterion-referenced evaluation)와 규준참조평가(norm-referenced evaluation)가 그것이다. 학생의 성취 결과를 미리 설정한 준거에 근거하여 해석하는 방식은 학생들이 무엇을 알고 할 수 있는지의 구체적인 성취 정보를 제공한다.

이론적인 분류 취지와는 달리, 현재 우리나라에서 시행되고 있는 평가는 초·중등학교 및 대학에서의 각종 평가 현장에서뿐만 아니라 기업체는 물론 정부부처, 공공기관 등 국가기관에서조차 준거참조평가(절대평가)의 방식만 도입하였을 뿐 실제 운영은 절대평가의 기본 개념조차도 적용되고 있지 않은 실정이다(경제·인문사회연구회, 2008).

다행스러운 것은 초등학교 3학년 국가수준의 기초학력진단평가(채선희 외, 2003)나 국가수준 학업성취도 평가(박정 외, 2006)와 같은 일부 국가시험에서나마 이러한 문제점에 대한 논의가 이루어지기 시작하였다는 것이다.

초·중등학교 학생들이 교실에서 사용하는 교과서에 대한 검정 심사의 적격·부적격 판정의 경우도 미리 합격본 수를 정해놓지 않고 심사를 시행하고 있는데, 앞서 언급한 평가의 두 가지 유형 중 준거 참조적 평가 방식에 해당된다고 하겠다.

이러한 교과서 검정에 대한 교과기준 항목별 심사에서 적격·부적격을 판정하는 준거는 제 7차 교육과정까지는 항목별 A, B, C 평정에 의한 C의 개수가 그것이었으나, 「2006년도 개정 교육과정에 의한 심사」부터 일반적으로 절대평가에서 정하는 수, 우, 미, 양, 가의 5가지 평가 단계 중 "우" 수준인 80점을 준거점수로 정하여 시행하고 있다.

따라서, 향후에는 심사본이 받은 점수와 항목별 검정기준 도달 여부를 가늠할 수 있는 준거참조평가의 철학과 원리가 반영된 분할점수를 다양한 방식으로 연구하여 제시할 필요가 있는데, 우선적으로 이러한 분할점수를 과학적으로 설정하여 실제 심사에서 적용을 한 후, 설정된 분할점수에 의하여 심사본들이 제대로 분류가 되었는지, 아니면 우연적인 요소가 가미되어 결과가 왜곡된 것은 없는지 분류일치도 추정을 통해서 반드시 점검을 해 보아야 할 필요가 있다.

검정 심사에서 쓰이는 '항목별 검정기준표'는 일반적인 수행평가 상황에서 하위 요인들의 구조를 나타내는 평가기준표가 되어, 교과서 검정에서 분할점수를 설정하고 분류일치도를 추정해 볼 수 있는 상황이 되는데, 검정 심사에서는 분할점수는 미리 일반적인 특정점수를 분할점수로 정하여 적격·부적격 판정을 내리고 있으므로 분할점수보다는 분류일치도 문제에 주목할 필요가 있다.

즉, 이렇게 정해진 분할점수를 활용하여 심사를 시행한 후에, 신뢰도 측면에서 심사 결과에 대한 분류일치도를 추정하여 2008년도에 시행된 교과서 검정 심사의 적격·부적격 판정의 수

준 분류가 일관성 있게 이루어졌는지를 분석하고, 만약 일치도 지수가 낮게 나온다면 분할점수가 제대로 설정되지 않아서 나온 문제인지 아니면 평가자들의 평가 자체가 문제인지를 분석하는데 본 연구의 목적이 있다.

Ⅱ. 이론적 배경

1. 교과서 검정 제도

우리나라의 교과서는 교육과정을 구현하기 위하여 학교 교육에서 교과교육의 실제를 안내하는 교과서, 지도서, 보조 자료를 의미하며 전통적으로 권위 있는 교육수단으로 학교 교육에서 수업은 교과서를 중심으로 전개된다. 우리나라의 교과서 제도는 국정, 검정, 인정으로 구분한다. 즉, 국가에서 발행 편찬하고 판권을 갖는 국정제와 일반 출판사에서 제작하고 편찬하여 일정한 검정기준에 맞는 교과서를 승인하는 검·인정제를 병행하고 있다.

검정기준은 대한민국 법질서, 교육과정 총론 및 편찬상의 유의점 등에 근거한 모든 교과에서 적용할 수 있는 보편적 기준을 심사하는 '공통기준'과 각 교과목별 특성에 부합하는 여부를 심사하는 '교과기준'으로 구성되어 있는데, 여기서는 분할점수 및 분류일치도와 직접 관련이 있는 '교과기준'에 대해서만 소개하기로 한다.

교과기준은 심사 영역과 심사 항목으로 구분하여 구성되어 있는데, 심사 영역은 '교육과정의 준수', '내용의 선정 및 조직', '창의성', '내용의 정확성 및 공정성', '교수·학습 방법 및 평가', '표기·표현 및 편집' 등 6개 영역으로 되어 있다.

교과기준에 의한 심사는 각 심사영역의 심사항목별 배점, 가중치, 심사관점을 교과용도서검정심의회에서 협의하여 결정하는데, 교과기준 상의 각 심사항목에 대하여 검정위원이 개별적으로 점수로 평정한다.

부적격 판정은 심의회의 심의결과 총점이 75점 미만이거나 1개 영역이라도 "60% 미만 점수"로 평정된 도서를 대상으로 하는데 이때의 총점은 각 심사위원이 평정한 총점의 합을 평균한 값을 말한다.

2. 분류일치도 추정

일반적인 평가 환경에서 피험자들을 미리 정해진 표준에 따라 몇 개의 겹치지 않는 범주로 분류하는 것은 일반적이다. 예를 들면, '숙달·비숙달' 또는 '통과·비통과'의 결정은 수십 년 동안

교육 분야에서뿐만 아니라 자격 및 인증 분야에서도 검사의 종합적인 역할을 해왔다.

최근 들어 한국교육과정평가원에서 수행하는 국가수준 학업성취도 평가에서는 다수의 범주(보통 4가지)로 분류하는 것에 주목하고 있다. 이러한 환경에서 아직도 신뢰도 등과 관련된 사항들에 역점을 두어야 함에도 불구하고 일반적인 사용자들은 검사점수의 일관성 그 자체보다도 분류 결정의 일관성에 중점을 두는 통계수치들을 항상 요구하고 있다. 형태나 상황에 따른 검사점수의 일관성은 중요하나 이러한 형태나 상황에 따른 분류 결정의 일관성과는 다르다.

따라서, 분류 일관성은 피험자들이 두 번 시행의 독립적인 동형 검사에서 일관성 있게 범주화되는 것으로 정의될 수 있다. 이러한 동형 검사는 두 가지 형태이거나 다른 두 번의 경우에 시행된 동일 검사지를 말한다.

두 가지 모두 검사 시행 시간이 두 배로 들며, 또 다른 검사에 필요한 요소들이 왜곡되는데 그래서 실제로 이러한 검사 시행은 어렵기 때문에 직접적으로 분류일치도 추정치를 구한다.

대부분의 분류일치도 추정 방식들은 〈표 1〉에서와 같이 다음의 두 가지의 구별되는 논리 유형을 따른다(Brennan & Wan, 2004; Lee, 2008a). 첫 번째 유형(Type I Logic)은 진점수 및 조건 오차원에 대한 분포를 추정하는데 직접적으로 집단에 대한 일치도를 추정한다.

반면 두 번째 유형(Type II Logic)은 개개인에 대한 조건 오차원에 대한 분포를 추정하는데 피험자 전체 집단에 대한 진점수 분포의 어떠한 가정도 없다. 이 유형은 한 번에 한 명의 피험자에 대한 일치도를 추정한 다음, 개별 피험자의 일관성 통계치를 평균값을 구하면서 전체 피험자 집단의 일치도 지수를 얻는다.

〈표 1〉 유형별 추정 방식 분류

유 형	특 징	방 식
유형 1 (Type I Logic)	분포를 가정, 전체 집단의 일관성 직접 추정	Huynh의 방식
		강진점수 모형
		L-L 방식
		Normal(NM) 방식
		B-L 방식
유형 2 (Type II Logic)	오차에 대한 조건 분포만 가정, 개개인 피험자의 분류 일관성 추정 후 전체 추정	Subkoviak 방식,
		Bootstrap 방식
		중다항(CM) 방식

"Standards for Educational and Psychological Testing(American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999)에 따르면, 검사 시행자는 분류 결정에 대한 신뢰도를

보고하는 것이 가장 중요한 책무이다"라고 하였는데, 최근의 검사 추세를 반영하여 볼 때, 이분 문항 보다는 복합 문항에 대한 분류 결정의 신뢰도에 우리의 관심사가 있다고 할 수 있다.

복합 문항에 대한 이러한 분류일치도 추정 방식들은 매우 다양한 방법들을 사용해서 분류일치도 추정 문제를 다루는데, 어느 것도 아직까지 그렇게 많이 연구되지는 못하였다. 이러한 방식들을 제안한 저자들은 보통 경험적인 자료들을 가지고 이러한 방식들의 유용성을 설명함에도 불구하고 논쟁은 그 방식들의 적절성에 대한 탐구보다는 종종 방법론의 안내에 국한되는 경우가 많다.

또한 저자들은 다양한 자료를 가지고 자기 방식에 적용도 거의 하지 않고 다른 방식과 자기 방식들을 비교도 하지 않았을 뿐만 아니라 사용자 친화적인 양질의 소프트웨어도 거의 없어 외부연구자들이 이러한 방식들을 제대로 연구하게 하지 못하였다.

이러한 문제의식을 가지고 문항반응이론에 근거한 방식을 사용하지 않는 추정 방식들을 소개하면 Normal-Approximation(NM) 방식, Breyer-Lewis 방식, Livingston-Lewis 방식, Bootstrap 방식, Compound-Multinomial(CM) 방식의 다섯 가지가 있다.

먼저, NM 방식을 소개하면, Huynh(1976)은 베타-이항 모형을 추정하면서 이분 문항에 대한 일치도 지수 p와 Kappa 계수를 추정하는 방식을 개발했는데, 상당히 큰 수의 문항들(n〉10)로 구성되는 검사에서는 이 방식을 이용하여 계산하는 것이 상당한 시간이 걸리는 일이 되었다.

시간의 문제를 해결하기 위하여 Huynh은 검사 점수 x를 변환된 형태로 정규 근사하는 아주 간단한 방법을 제안했는데, Peng & Subkoviak(1980)은 이러한 정규근사 방법을 훨씬 더 간단하게 하였고 Huynh의 방식보다 더 좋다는 증거를 제시했다.

이러한 방식은 베타-이항 분포의 가정이 필요없게 하였고 대신 상관계수로 KR-21 사용과 함께 두 동형검사로부터 얻은 점수에 대한 이변량 정규분포를 가정한다. 이러한 가정을 사용하게 되면, 의사결정의 일관성 지수는 아주 쉽게 계산될 수 있다.

NM 방식은 프로그램에 의한 결과를 바로 구할 수 있는 방식이 아니므로 추정치를 구하는 절차를 간단히 제시해보면 다음과 같다.

첫째, 일관성 있는 의사결정의 비율인 P를 구하는 개념은 〈표 2〉에 제시하였다.

〈표 2〉 일관성 있는 의사결정의 비율

Y형 검사 / X형 검사	실패	성공	주변치
실패	$P_{00}(A)$	$P_{01}(B)$	$P_{0\cdot}$
성공	$P_{10}(C)$	$P_{11}(D)$	$P_{1\cdot}$
주변치	$P_{\cdot 0}$	$P_{\cdot 1}$	1.0

둘째, 평균과 표준편차, 분할점수를 이용하여 z값을 구하고, 신뢰도 계수를 확보한다. 셋째, 이변량 정규분포의 개념을 적용하여 STATDIST 프로그램(아이오와대학의 공유 프로그램)에서 확률값을 구한다. 이러한 절차를 거쳐 위에서 안내한 식에 수치를 넣으면 P와 Kappa 계수를 구할 수 있다.

Breyer-Lewis(1994) 방식은 하나의 검사를 두 부분의 검사로 나누는 것이 필요한데 비교 가능해야 하지만 반드시 동형 검사일 필요는 없으며 각 부분은 분할점수를 갖고 있으며 두 부분의 분할점수 합은 전체 검사의 분할점수와 같아야 한다. 이러한 검사는 반분 검사의 분할표에 따른 이변량 정규분포가 기본적으로 가정되어야 하는데 반분 검사 점수의 이변량 정규분포의 가정보다는 약한 가정이라고 주장한다. 그렇다면 사실은 이른바 '이변량 정상성이 기본' 이라는 가정은 희미해진다.

이 방식은 먼저 반분 검사에 대하여 4개의 상관관계를 추정하고 그 것을 반분 검사의 효과적인 신뢰도의 추정치로 여긴다. 그런 다음 전체 검사에 대한 신뢰도를 추정하기 위하여 Spearman-Brown 예측 공식을 적용한 다음, 역으로 전체 검사에 대하여 적용된 이변량 정규분포를 위한 상관계수로서 그 추정치를 사용한다.

위의 결과가 주어진 상태에서 전체 검사의 의사결정의 일관성은 다음과 같은 계산 과정을 거쳐 추정될 수 있다. BL 방식과 NM 방식 모두 이변량 정상성 가정에 기본을 두는데 두 방식은 각각 다소 다른 방법으로 그 가정을 사용한다. NM 방식은 직접적이고 전적으로 그 가정의 기초 하에서 개발되었는데 논의의 방향이 분명한 반면, BL 방식에 적용된 이변량 정상성 가정은 직접적이지 못하고 논의의 방향도 다소 특별하게 나타난다.

Breyer & Lewis(1994)는 그들의 방식이 '간단한' 경우에서의 Subkoviak 방식(1976)과 복합적인 경우에서의 LL 방식(1995)에서 유사한 결과들을 얻었다는 것을 보여 주었는데 BL 방식은 적절히 연구되지 못했기 때문에 그 방식의 유용성을 증명하기 위해서는 추가적인 연구가 필요하다.

다음으로, Livingston-Lewis 방식(1995)은 복합 문항을 다루려고 하는데 Hanson & Brennan(1990)에 의해 제안된 강진점수 방법과 공통적인 요소를 많이 가지고 있다. 예를 들면, Livingston-Lewis 방식에서 진점수 분포는 4모수 베타 분포의 적합성에 의하여 추정되고 오차의 조건분포는 이항분포의 적합성에 의하여 추정된다.

이 방식의 핵심은 Livingston & Lewis가 복합 자료 모형을 만들기 위한 이른바 "유효 검사 길이($\tilde{n}$으로 표기)"를 고안한 것이다. 이 용어는 실제 사용된 점수를 가지고 동일한 정확도를 갖는 전체 점수를 산출하는데 필요한 이산적인, 이분 문항의, 지역적 독립성을 가진 검사 문항의 수를 말한다. 유효 검사 길이를 산출하는 공식은 다음과 같다.

$$\tilde{n}= \frac{(\mu_x - X_{\min})(X_{\max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)}$$

$\tilde{n}$: 반올림 처리된 정수
$X_{\min}$: 최저 점수
$X_{\max}$: 최고 점수
μ_x : 평균, σ_x^2 : 분산, r : 검사 신뢰도

여기에서 X는 보고된 점수를 말하는데, 정답을 한 문항의 수 또는 습득한 점수의 수라는 의미로 볼 때 이 점수는 반드시 원점수일 필요는 없다. 그렇기 때문에 Livingston-Lewis 방식은 분류일치도를 추정하는데 척도 점수를 사용할 수 있다.

Livingston & Lewis가 연구한 여러 시험들에 대하여 이 방식은 높은 정확도를 보여 줬다. 그들은 각각의 시험들을 두 개의 동등한 부분으로 나눠서 추정치를 도출했는데, Livingston-Lewis 방식으로 얻은 P추정치와 직접 두 개로 나뉜 검사를 비교하여 얻은 추정치 사이의 차이는 보통 약 0.01에 불과하였다.

BW 방식(Bootstrap Procedure)의 Bootstrap은 원래 Efron(1982)에 의하여 개발된 재표집 알고리즘으로 모집단 모수의 특정 추정치의 정확도를 평가하기 위한 것이다. Bootstrap 알고리즘은 본래의 표집을 다수의 임의 표집으로 교체하는 것을 포함한다. 분류일치도를 추정하기 위한 Bootstrap 방식은 문항 반응 벡터의 Bootstrap 표집을 생성하는 것으로 시작한다.

그런 다음, 본래의 자료에 사용된 채점/척도화 규칙을 Bootstrap 표집에 적용한다. 각각의 피험자에 대하여, 그 피험자가 본래의 형태에서와 마찬가지로 반복된 결과에서도 동일한 성취범주로 분류될 때, 일관성 있는 의사 결정이 내려진다. 이러한 과정은 수없이 반복된다.

개별 피험자에 대하여, 모든 반복 시행 결과를 통해 일관성 있는 의사 결정의 비율이 계산될 수 있는데 이것은 개별적인 일관성 지수 P를 의미한다. 모든 피험자들의 개별 P값을 평균하면 전체에 대한 P값을 추정할 수 있다.

위에 기술된 대로 이 방식은 차별적이지 않은 문항들로 구성된 단순한 검사를 위한 방식인데, 그러한 단순한 검사에서는 의사결정의 일관성을 계산하기 위하여 Bootstrap 방식을 꼭 써야 할 특별한 이유는 없다. 왜냐하면 Huynh이나 강진점수 방식 같은 다른 방법들이 직접적으로 사용될 수 있기 때문이다.

이와는 달리, 복합 문항에서 분포함수들의 직접적인 이용이 어려울 때, Bootstrap 방식을 적용하는 것이 상대적으로 간단하다. 복잡한 상황에서는 층화된 Bootstrap 표집 방식을 통해서 평가를 반복하는 것이 필요하다는 것을 알아야 한다.

Brennan & Wan(2004)은 Bootstrap 방식에 대하여 두 가지 접근 방식에 의하여 결정되는 의사결정의 일관성을 명시적으로 특징 지웠다. 그러나 이러한 특징은 특별히 이 방식에만 국한되지는 않는데, 대신에 분류 일관성을 추정하는 어떠한 맥락에도 적용될 수 있다.

하나의 방법은 "의사결정의 일관성은 두 번의 반복 시행된 검사의 토대에서 결정된다"는 것이다. 그 형태는 Bootstrap 방식의 문항 재표집을 통해서, 또는 Livingston-Lewis 방식의 미리 추정된 분포 모형(베타-이항 분포)을 통해서 생성될 수 있다. 이러한 두 가지 가상 형태로부터 얻은 검사 결과는 분류 일관성을 결정하는데 고려된다.

이러한 방법은 여러 연구물에서 보통 일반적으로 보여 지는데, 그래서 앞부분의 분류일관성 지수의 유도에서는 주로 이러한 방법에 대하여 설명하고 있다. 다른 방법은 "의사결정의 일관성은 실제 시행된 검사와 반복하여 시행된 검사의 두 형태를 기반으로 결정된다"는 것이다.

Brennan & Wan(2004)은 실제 검사에 기반을 둔 검사 결과가 고려될 필요가 있다고 주장했는데, 왜냐하면 특정한 검사에서 합격/불합격을 받은 피험자에 대한 조작적인 상황에서, 만들어 질 수 있는 유일하고 일관성 있는 의사결정은 반복 시행에서도 상응하는 '합격·불합격' 결정이 나오는 것이다.

마지막으로, CM 방식(Lee, 2008a)은 비차별적인 다분 문항으로 구성된 검사에 대하여 다항 오차 모형을, 서로 다른 문항 세트의 조합으로 구성된 검사에 대하여 중다항 오차 모형을 사용한다. 또한 의사 결정의 일관성을 추정하는 것과 함께 개별적인 측정의 표준오차 및 복합 문항에 대한 신뢰도도 추정한다(Lee, 2001, 2008a).

다항 오차 모형 방식은 유형 2(Type Ⅱ Logic)에 속하는데, 단지 두 개의 점수대를 가지고 있을 때는 Subkoviak 방식(1976)으로 환원된다. $g_1 < g_2 \cdots < g_h$와 같은 $h(h>2)$개의 점수 범주를 가진 n개의 다분 문항을 포함하는 검사를 가정해 보자. 이 문항들은 차별화되지 않은 임의의 문항들로 구성되었다고 추정된다. $\tau = \{\tau_1, \tau_2, \cdots, \tau_h\}$는 각 피험자가 각각 $g_1, g_2, \cdots, g_h$의 점수를 얻을 수 있는 전집에서의 문항들의 비율을 말한다. 또한, $X_1, X_2, \cdots, X_h$는 각각의 점수대에서 개별적으로 점수를 획득한 관찰 문항들의 수를 나타내는 임의변수들인데 이 변수들의 합은 $n(X_1 + X_2 + \cdots + X_h = n)$이며 각각의 개별적인 피험자들에게 이러한 임의 변수들은 다항분포를 따른다.

$$\Pr(X_1 = x_1, X_2 = x_2, \cdots, X_h = x_h \mid \vec{\tau}) = \frac{n!}{x_1! x_2! \cdots x_h!} \overrightarrow{\tau_1^{x_1}} \overrightarrow{\tau_2^{x_2}} \cdots \overrightarrow{\tau_h^{x_h}}$$

$\tau = \{\tau_1, \tau_2, \cdots, \tau_h\}$는 동일한 점수를 얻은 문항들의 비율에 의하여 추정될 수 있다. $x_1, x_2, \cdots, x_h$ 값들의 서로 다른 많은 세트들은 특정한 총점 y로 이끌 수 있는 가능성이 있으

므로 Y의 확률밀도함수(Probability Density Function; PDF)는 총점 y를 산출하는 $x_1, x_2, \cdots, x_h$의 모든 세트들을 더하는 것으로 얻을 수 있다.

$$\Pr(Y=y \mid \vec{\tau}) = \sum_{g_1x_1+g_2x_2+\cdots+g_hx_h=y} \Pr(X_1=x_1, X_2=x_2, \cdots, X_h=x_h \mid \vec{\tau})$$

Ⅲ. 연구 방법

1. 연구 자료

교과서 검정 심사의 적격·부적격에 대한 점수제 판정 결과의 분석을 위하여 〈표 3〉과 같이 「2006년 개정 교육과정에 의한 수학·영어 교과서 검정」 심사 자료를 활용하였다. 이러한 검정 심사는 현장 교사들로 구성된 연구위원 220명이 수행한 기초조사와 현장 교사와 대학 교수들로 구성된 118명의 검정위원들이 수행한 1차, 2차 검정 심사로 구성되어 시행되었다.

먼저, 이 논문에서 검정 심사 자료 분석에 의한 연구(분류일치도 추정 연구)는 1차 심사를 수행한 검정위원들이 각 심사본의 검정 기준상 항목들에 대하여 개별적으로 평가하고 채점한 점수 자료를 가지고 연구가 이루어졌다.

〈표 3〉 검정 대상 도서(총 8종) 및 출원 책 수

학 교 급		교과서 명	책 수	검정위원 수
중학교	1학년	수학1, 수학 익힘책1,	각 42책	33명
		영어1, 영어1 학습활동책	〃	23명
고등학교	1학년	수학, 수학 익힘책	각 38책	41명
		영어, 영어 학습활동책	각 30책	21명

일반적인 검사 시행상의 문항에 해당하는 검정기준 상의 심사항목 수는 수학은 19개(익힘책은 20개), 영어는 25개로 영어가 다소 많은 수준이며, 배점은 학교급 및 과목별, 책별로 약간씩 다른 1점에서 15점 사이로 배점 단계의 수는 5개에서 7개까지였다. 직접적인 분석의 대상이 되는 자료의 수(검정위원수×책수)는 수학이 중학교 1,386개, 고등학교 1,558개로 영어의 중학교 966개, 고등학교 630개보다 많았다.

각 항목의 배점에 대한 점수는 수, 우, 미, 양, 가의 5단계 척도로 100%부터 75%, 50%,

25%, 0%까지 부여하여 채점을 하였는데, 예를 들어 어느 항목의 배점이 4점이고 평정자가 우를 평정하였다면 전산프로그램에서는 자동적으로 3점이 채점되게 하였다.

한편, 교과서 검정 심사 판정의 종류는 적격(80점 이상), 판정유예(75점 이상 80점 미만), 부적격(75점 미만)의 세 가지로 구분되나 실질적으로 판정유예의 경우 적격 판정과 구별되지 않으므로 적격으로 분류하기로 하였다.

〈표 4〉 검정 대상 도서별 평가항목 및 배점

구 분			항목 수 (검정기준)	배점		비 고
				개수	범위	
중학교	수학	본책	19	7	3~14	
		익힘책	20	7	3~14	수학익힘책
	영어	본책	25	5	2~15	
		활동책	25	6	2~15	학습활동책
고등학교	수학	본책	19	6	3~14	
		익힘책	20	6	2~14	수학익힘책
	영어	본책	25	6	1~15	
		활동책	25	7	1~15	학습활동책

또한, 심사본의 부적격 처리 기준은 '총점 기준 75점 미만' 이외에 '영역별로 60% 미만 점수가 있을 경우'의 한 가지가 더 있으나 부적격본들 대부분이 '총점 기준 75점 미만'에 해당되어 부적격 처리 되었는데, 중학교 1학년 수학 익힘책에서만 전체 총점은 75점 이상으로 적격 기준에 해당하고 '영역별 60% 미만 점수' 영역이 있는 심사본이 일부(3권) 있었으나 전체 자료 분석에 미치는 영향이 미미하여 이러한 기준을 전체 자료 분석에는 고려하지 않았다.

2. 판정 결과의 분류일치도 추정

복합 문항들에 대한 분류일치도 추정은 Wan(2008)이 제시한 다섯 가지 비-IRT 방식을 이론적 배경에서 살펴본 바 있는데, 이 논문에서는 유형 1(Type Ⅰ)에 속하는 NM 방식과 Livingston-Lewis 방식 두 가지를, 유형 2(Type Ⅱ)에 속하는 방식 중 CM 방식 한 가지를 선택하여 분석하였다.

먼저, NM 방식에서 신뢰도 지수는 NM 방식으로 분류일치도를 추정하는데 가장 중요한 요소인데, Peng & Subkoviak(1980)은 이분 문항의 경우에는 $KR-21$ 을 쓰도록 제안했으며

복합 문항 자료에는 층화 계수 알파나 Feldt-Raju 지수, 일반화가능도 이론에서의 의존도 계수(Φ)같은 내적 일관성(신뢰도) 지수를 얻는 여러 가지 방법들이 있다고 하였다.

의사 결정의 맥락에서는 다른 피험자와의 상대적인 점수 비교보다는 피험자 점수의 절대적 기준이 더욱 관심이 되는 사항으로 위의 여러 가지 절대평가 관련 지수 중 적용 가능한 지수로 일반화가능도 이론(Brennan, 2001)의 $\sigma^2(\Delta)$ (절대오차 분산)과 관계된 계수인 Φ(phi) 수치를 선택하여 신뢰도 계수로 활용하여야 한다. 이 Φ(phi) 계수는 다변량(multivariate) 일반화가능도 이론에 의하여 추정되어야 하나 검정 심사의 평정에서는 검정기준의 점수 단계가 많고 단계마다의 점수 차도 매우 크면서 단일 문항으로 구성된 점수대도 있어 다변량 일반화가능도 이론 설계에 의한 Φ(phi) 계수를 산출하기가 곤란하여 α계수를 사용하였다.

다만, 이 α 계수는 Lee(2007)가 제안한 "CM 방식으로 구한 신뢰도 값과 Φ값이 유사"하다고 한 결과를 활용하여 고등학교 수학 본책 및 익힘책에서 단일 문항을 제거한 상태로 구한 α 계수와 마찬가지로 동일 책에서 단일 문항을 제외하고 구한 CM 방식의 내적일치도 수치를 가지고 Livingston-Lewis 방식에 적용하여 그 분류일치도 추정치를 비교하여 이 α계수를 NM과 Livingston-Lewis 방식의 신뢰도 계수로 활용할 수 있는가를 확인하였다.

이러한 신뢰도 추정치가 있는 상태에서 분할점수는 한 개이므로 P와 p_c는 다음과 같이 간단하게 설명될 수 있는데, STATDIST 프로그램을 이용하여 단변량 및 다변량 정규분포의 누적확률밀도(CDF)를 계산하여 NM의 추정치를 구하였다.

$$P = 2[p(x < z,\ y < z) - p(x, z)] + 1$$

$$p_c = [p(x < z)]^2 + [1 - p(x < z)]^2$$

z : 표준화된 분할점수

다음으로, Livingston-Lewis 방식은 다음의 네 가지 요건을 필요로 하는데 첫째, 실제 시행된 검사의 점수 분포 둘째, 점수에 대한 신뢰도 계수 셋째, 검사에서의 최댓값 및 최솟값 넷째, 분할점수가 그것으로, 셋째의 신뢰도 계수는 일반화가능도 이론 방식(Brennan, 2001)이나 CM 방식(Lee, 2007)으로 구할 수 있다.

CM 방식에 의한 신뢰도 계수는 이 논문에서는 특정 배점에 속하는 문항이 한 개에 불과한 경우가 여러 개가 있어 오차 계산에서 작용을 하지 못하여 신뢰도 계수 수치가 크게 나오는 경향이 있으므로 NM 방식에서와 같이 동일한 α계수를 신뢰도 계수로 사용하였다.

Livingston-Lewis 방식에서는 BB-CLASS 프로그램(Brennan, 2004)이 지수 계산에 사용되었는데 이 프로그램은 Hanson & Brennan(1990)에 의한 강진점수 방식과 Livingston-

Lewis(1995) 방식의 분류일치도 값을 제공하기 위하여 개발되었다.

마지막으로, CM 방식은 MULT-CLASS(Lee, 2008b) 컴퓨터 프로그램을 사용하여 수행되었는데, 자료 조건이 정수이므로 전체 분석 경향에 영향을 주지 않아 소수점 첫째 자리에서 반올림한 자료를 가지고 분석하였다.

앞부분에서 여러 번 언급된 대로, CM 방식은 이전의 NM 방식 및 Livingston-Lewis 방식과는 달리 전체 검사를 대상으로 하는 분포가 아닌 개개인에 대한 조건 오차원에 대한 분포를 추정하는데 피험자 전체 집단에 대한 진점수 분포의 어떠한 가정도 없다. 이 유형은 한 번에 한 명의 피험자에 대한 일치도를 추정한 다음 개별 피험자의 일관성 통계치의 평균값을 구하면서 전체 피험자 집단의 일치도 지수를 얻는다.

이 논문에서는 학교급 및 과목별로 배점의 종류가 5~7개로 상당히 많아 프로그램 실행 시간이 많이 걸렸는데, 유형이 같고 배점 종류가 3개 이하인 자료는 즉시 결과 산출이 가능한 것과 비교하여 대조가 되었다.

3. 모형 적용을 위한 가정 적합성

이 논문에서 연구된 분류일치도 추정 방식들은 모두 분포 형태의 가정을 기반으로 해서 추정되었다. 즉, NM 방식은 단변량 또는 이변량 정상성 가정에 따랐고 Livingston-Lewis 방식은 4모수 베타-이항 모형을 사용하였다. 이러한 가정들이 실제 관찰점수 분포와 잘 맞는지를 검증하는 것이 이 추정 방식들의 유용성을 평가하는 길이 될 것인데, 이 논문에서는 전체 자료에 대한 별도의 분포, 즉 주변 관찰점수 분포에 대한 별다른 관심이 없고 또 분포 가정을 확인하는데 한계가 있는 CM 방식을 제외하고 나머지 두 가지 즉, NM과 Livingston- Lewis 방식의 모형 적합성을 점검하였다.

먼저, NM 방식에서 단변량이나 이변량 정상성 가정은 NM 방식의 가장 중요한 가정이기는 하나, 표본 크기가 크고(20개 이상) 분석이 평균값에 의존하는 상황에서는 정상성 가정이 그렇게 중요하지는 않다. 다만, 어느 정도는 통계적 기법을 사용하는 추론의 질은 모집단이 다변량 정상 형태에 얼마나 가깝게 닮아 있나에 달려 있다.

또한 낮은 차원의 정상 분포(normal distribution) 상태가 더 높은 차원에서 비정상 분포 상태는 쉽게 일어나지 않는다(Johnson & Wichern, 2002). 즉, 단변량의 경우 정상성을 갖도록 하면 완전하지는 않지만 이변량의 경우도 정상성의 문제를 해결할 수 있다는 의미이다.

NM 방식의 모형 적합성 검증은 단변량 정상성에 대하여 이러한 가정이 실제 시행 자료에도 잘 적용되는지를 Q-Q plot을 그려 검증하였다. 이변량 정상성 가정은 chi-square plot을 그려 확인하는데, 대규모 단일 검사의 경우에서는 동형의 반분검사 세트가 기본적으로 가정되어야

하나 여기서는 문항(검정 항목) 사이의 점수차가 아주 크고 배점이 다양한 점수대로 구성되어 있어 동형의 반분검사로 구분하는 것이 커다란 오차를 가져오게 되어 별도로 시행하지 않았다.

Q-Q plot은 각 심사본에 대하여 주변 정상성 가정의 성립 여부를 보여주는 직선 형태를 확인한다. 이러한 plot은 표본 변위치 대 관찰된 자료들이 정말로 정규분포 형태를 띤다면 보여질 수 있는 변위치(quantile)를 그리는데 직선에 가까우면 이러한 정상성 가정을 인정할 수 있다.

Livingston-Lewis 방식의 분석에서는 실제 두 개의 충분한 검사 길이를 가진 검사 형태가 없어 보통의 자료들은 단일 시행 검사에서 나오기 때문에 이러한 가정들을 확인할 방법이 없는데, 충분한 검사 길이와 거의 동형의 반분 검사들을 활용하여 동일한 효과를 얻도록 하였다. 또한, 4모수 베타-이항 모형은 Livingston-Lewis 방식 추정의 필수 부분인데 실제 심사 시행 자료의 적합성을 검증하기 위하여 효과적인 검사 길이에 의한 실제 자료와 모형의 자료에 대한 점수 분포도를 그려 비교하였다.

Ⅳ. 연구 결과

이 논문에서는 2008년도에 시행된「2006년도 개정 교육과정에 따른 수학·영어 교과서 검정」심사의 결과 자료를 가지고 기본적인 기술 통계량을 도출하였으며, 비-IRT 방법의 세 가지 분류일치도 추정 방식으로 구한 P계수 값과 Kappa 계수 값을 비교·분석하였다.

1. 심사 결과의 기술 통계량

전체적인 평균값은 학교급 및 책별로 임의로 정해진 분할점수 75.0점보다 높은 75.8점에서 89.0점까지로 대부분의 책에서 부적편포를 보이는데, 고등학교 영어 및 영어 학습활동책에서는 정적편포를 보이고 있다.

또한, NM 방식의 분류일치도 추정과 Livingston-Lewis 방식의 추정에서 가장 중요한 역할을 하는 신뢰도 지수(α계수)는 고등학교 수학 본책과 익힘책의 0.93과 같이 대부분의 책에서 0.68 이상의 높은 수치를 보이고 있어 일반적인 기준에 비추어 보더라도 전체적인 평정 점수의 신뢰도는 양호하다고 할 수 있다.

다만, 이 α계수는 연구방법에서 설명한 대로 Lee(2007)가 제안한 "CM 방식으로 구한 신뢰도 값과 Φ값이 유사"하다고 한 결과를 활용하였다. 고등학교 수학 본책 및 익힘책에서 단일 문항을 제거한 상태로 구한 α계수와 마찬가지로 동일 책에서 단일 문항을 제외하고 구한 CM 방

식의 내적일치도 수치를 가지고 Livingston-Lewis 방식에 적용하여 그 분류일치도 추정치를 비교한 결과, 큰 차이가 없어 이 α계수를 Φ값 대신 NM과 Livingston-Lewis 방식의 신뢰도 계수로 활용하였다.

〈표 5〉 학교급 및 책별 기술 통계량

구 분			평정 점수				표준 편차	왜도 (skew)	첨도	α 계수
			평균	최솟값 (A)	최댓값 (B)	차 (B-A)				
중학교	수학	본책	81.7	53.8	97.0	43.2	9.2	-0.6	2.3	0.84
		익힘책	89.0	67.3	99.0	31.7	6.6	-1.0	3.3	0.68
	영어	본책	78.5	55.0	96.5	41.5	8.8	-0.4	2.2	0.75
		활동책	75.8	54.3	93.0	38.7	10.0	-0.1	1.6	0.74
고등 학교	수학	본책	79.7	48.5	100.0	51.5	13.3	-0.2	1.5	0.93
		익힘책	79.3	50.0	100.0	50.0	15.1	-0.3	1.5	0.93
	영어	본책	80.7	67.3	96.5	29.2	7.1	0.1	2.1	0.79
		활동책	76.7	66.5	93.5	27.0	7.3	0.5	2.5	0.75

2. 분류일치도 분석 결과

개정 교육과정에서 적용한 분할점수는 사전 연구에 의한 점수가 아닌 일반적으로 학교 현장에서 적용하는 우(80점) 수준의 점수를 반영한 것으로, 여기에 판정유예 판정 변수를 고려하여 우(80점)와 미(70점)의 중간 수준인 75점으로 점수를 조정한 것이므로 학교급 및 책별 분할점수는 모두 75점인 동일한 조건으로 분석하였다.

NM 방식과 Livingston-Lewis 방식, CM 방식의 세 가지 추정 방식에 따른 일치도 지수(P) 수치는 〈표 6〉에서와 같이 모든 책에서 거의 차이가 없을 정도로 비슷한 경향성을 보였으며, 일치도 지수(P)값은 대부분 0.7~0.9 정도로 매우 높은 수준을 보였다.

다만, CM 방식의 추정치는 전체 책에서 타 방식에 비하여 다소 높게 나왔는데 이러한 현상은 "'개별 심사본의 항목별 점수를 모두 대상으로 하여 분석하는 추정 방식의 차이'와 그에 따라 '동일 배점을 가진 항목 수가 한 개인 경우에 누락되는 개별 오차의 차이'에서 비롯되었을지도 모른다"고 해석할 수도 있다.

성태제(1989)는 채점자간 신뢰도 추정으로 채점 자료에 대한 신뢰성을 인정하는 절대적 기준은 없으나 채점 결과가 점수로 부여될 때, 상관계수가 0.6 이상, 그리고 채점 결과가 범주로

부여될 때 일치도 지수(p)는 0.85 이상, Kappa 계수는 0.75 이상을 제안하고 있다.

이러한 기준으로 판단하였을 때, NM 방식과 Livingston-Lewis 방식의 영어 학습활동책을 제외한 대부분에서 일치도 지수(P)는 높은 수준의 매우 양호한 분류일치도를 보였음을 알 수 있다.

〈표 6〉 추정방식 및 책별 일치도 지수(P) 비교

구 분			분할 점수	추정 방식		
				NM	Livingston -Lewis	CM
중학교	수학	본책	75	0.86	0.89	0.91
		익힘책	75	0.97	0.93	0.97
	영어	본책	75	0.78	0.85	0.91
		활동책	75	0.76	0.78	0.93
고등학교	수학	본책	75	0.89	0.94	0.97
		익힘책	75	0.88	0.96	0.98
	영어	본책	75	0.84	0.78	0.86
		활동책	75	0.77	0.72	0.86

Kappa 계수는 〈표 7〉에서 보면 중학교 수학 본책과 고등학교 수학 본책 및 익힘책에서 일반적인 기준상 높은 수준의 추정치를 보였으며, 나머지 책들에서는 CM 방식을 제외한 NM 방식과 Livingston-Lewis 방식에서 다소 낮은 수준의 추정치를 보였는데 전반적으로 일치도 지수(P)보다 분류 일치 정도가 다소 낮게 나온 것을 알 수 있다. 즉, 일치도 지수(P)보다 우연에 의한 값을 제외하고 산출한 Kappa 계수가 낮게 나왔다는 것은 일반적인 분류일치 정도보다 우연에 의한 요소를 배제한, 진정한 분류 일치 정도가 다소 좋지 않다는 것을 의미한다.

CM 방식의 경우는 일치도 지수(P)보다 수치는 다소 낮지만 대부분 기준 수치를 상회하는 정도로 앞의 두 방식보다는 월등히 높은 일치 정도가 나온 것을 알 수 있다.

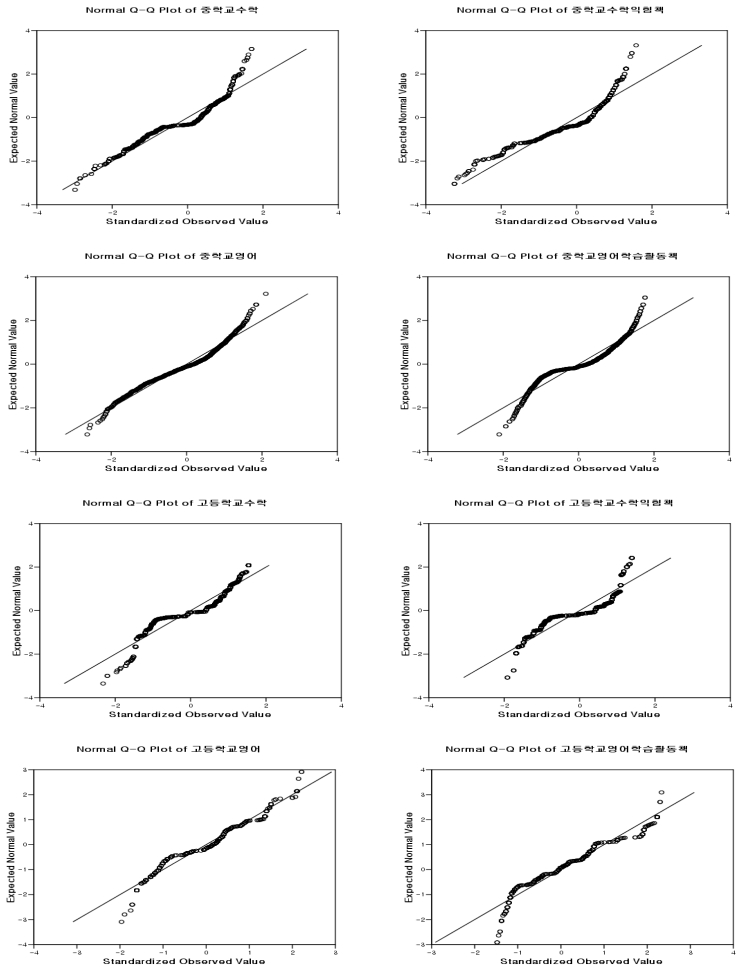
다만, 고등학교의 수학 본책 및 익힘책에서의 Kappa 계수는 일치도 지수와 함께 세 가지 방식 모두에서 다른 책들보다 월등히 높은 수치가 나왔는데 이러한 수치가 실제 분류가 적절하게 이루어진 결과인지, 아니면 평정 당시 독립적인 판단에 의한 것이 아닌 담합에 의한 사전 분류 후에 기계적으로 점수를 부여한 것에 따른 결과인지 쉽게 확인할 수 없는 문제로 전문가에 의한 심사 상황에서 나타나는 특이한 현상으로 볼 수 있다.

〈표 7〉 추정방식 및 책별 Kappa 계수(κ) 비교

구 분			분할 점수	추정 방식		
				NM	Livingston –Lewis	CM
중학교	수학	본책	75	0.61	0.72	0.73
		익힘책	75	0.30	0.15	0.62
	영어	본책	75	0.53	0.67	0.79
		활동책	75	0.52	0.55	0.86
고등학교	수학	본책	75	0.76	0.87	0.93
		익힘책	75	0.76	0.92	0.96
	영어	본책	75	0.55	0.39	0.64
		활동책	75	0.54	0.44	0.62

3. 모형 적용을 위한 가정 적합성 검토

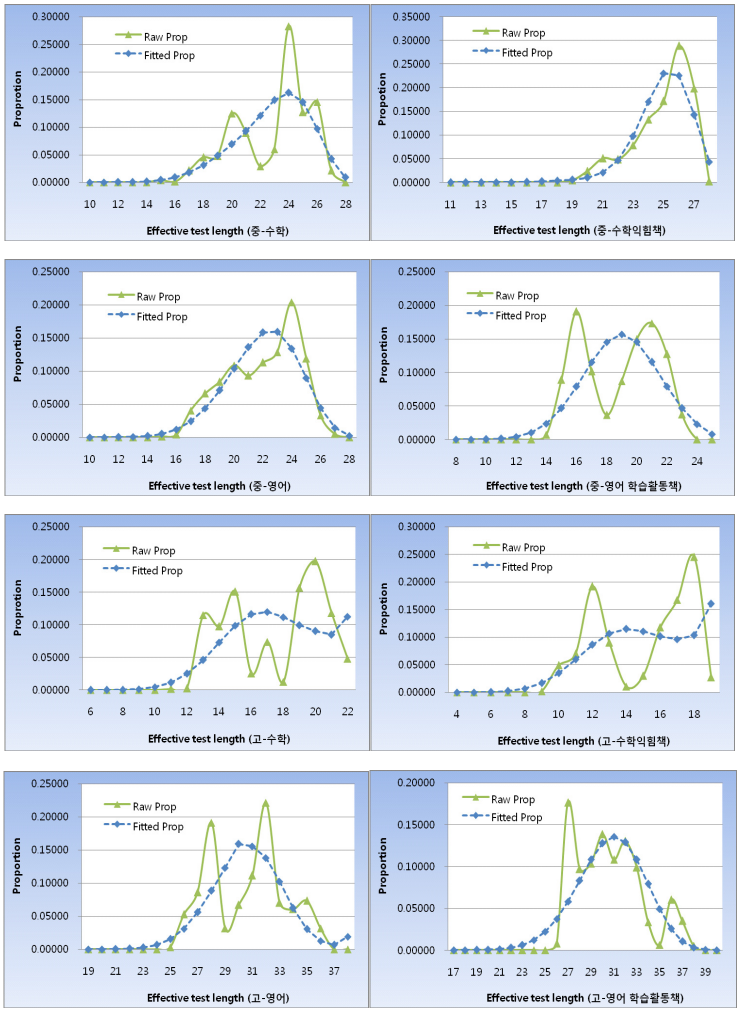
먼저, 적합성 검토를 위하여 NM 방식의 정상성 가정을 확인해 보면, 다음의 〔그림 1〕의 Q-Q plot은 중학교 및 고등학교 각 교과서의 검정위원별 평정 결과인 총점에 대한 직선 여부로 정상성 가정이 성립하는가를 보여 주는데 대부분의 심사본에서 직선의 모습을 보여 일단 정상성은 성립하는 것으로 보이나 이상값(outlier)이 책별로 다소 나타나는 것을 알 수 있다.



〔그림 1〕 학교급 및 책별 Q-Q plot

또한, Livingston-Lewis 방식 추정에서 4모수 베타-이항 모형 가정에 대한 실제 심사 시행 자료의 적합성을 검증하기 위하여 〔그림 2〕에서와 같이 유효 검사 길이에 의한 실제 자료와 모형의 자료에 대한 점수 분포도를 그려 비교하였다.

일반적으로 대부분의 책들에서는 시행 자료와 모형의 분포도는 일치하는 모습을 보였으나 앞의 분류일치도 결과와는 달리 고등학교 수학 본책 및 익힘책의 분포도는 잘 맞지 않는 모습을 보여 기술 통계 수치에서의 부조화와 함께 충분히 검토되어야 할 사항으로 여겨진다.



〔그림 2〕 학교급 및 책별 베타-이항 모형의 점수 분포도

Ⅴ. 결론 및 제언

1. 요약 및 결론

교과서 검정 심사에서 도서별로 해석되는 수준의 특성은 교육 내용의 오류가 없고 이념적으로 편향되지 아니한 교과서로서 교육과정이 추구하는 인간상과 교육 목표 달성에 적합한 질 높

은 교과서인가와 교육과정 중심의 학교 교육 체제에 적합한 학습자 중심의 질 높은 교과서인지를 심사하는데, 이러한 심사본이 검정 심사 기준을 어떻게 충족시켰는가를 평가한다. 이러한 평가가 가능하려면 점수척도를 두 부분으로 나누는 분할점수 하나를 설정해야 하는데, 이러한 분할점수의 설정은 실제 검사에서 적용을 한 후 설정된 분할점수에 의하여 피험자들이 제대로 분류가 되었는지, 아니면 우연적인 요소가 가미되어 결과가 왜곡된 것은 없는지를 반드시 점검을 해 보아야 한다.

Cizek(2001)은 기준을 설정하는 것은 많은 직업 분야에서 확실하고 효과적인 방법들을 발전시킨다는 점에서 이익을 주고 공공의 이익을 보호한다는 점에서 중요한 기능을 수행한다고 하였다. 그런데 우리나라에서는 검사에서 이렇게 중요한 기능을 하는 기준 설정 자체도 일반화되지 못하고 시행되지 않는 상태에서 검사가 아닌 교과서 검정에서 분류된 결과에 대한 분류일치도 추정은 더더욱 쉽지 않은 문제이다.

이러한 현실을 바탕으로 지금까지 일반적으로 연구되어 왔던 검사문항에 대한 분류일치도 추정이 아닌, 교과서 검정 심사본의 적격, 부적격 판정 결과에 대한 분류일치도 추정을 시행해 보았다.

교과서 검정 심사 결과 검정위원별 평정 자료의 전체적인 평균값은 학교급 및 책별로 분할점수 75점보다 높은 75.8점에서 89.0점까지로 대부분의 책에서 부적편포를 보인 반면 고등학교 영어 및 영어 학습활동책에서는 정적편포를 보이고 있다. 학교급 및 책별 평정 점수 분포는 대부분의 교과에서 분할점수 75점의 점수대를 중심으로 퍼져 있는 양상을 보인다.

또한, NM 방식의 분류일치도 추정과 Livingston-Lewis 방식의 추정에서 가장 중요한 역할을 하는 신뢰도 지수는 고등학교 수학 본책과 익힘책의 0.93과 같이 대부분의 책에서 0.68 이상의 높은 수치를 보이고 있어 일반적인 기준에 비추어 보더라도 전체적인 평정 점수의 신뢰도는 양호하다고 할 수 있다.

다만, 고등학교 수학 과목의 상대적인 높은 신뢰도는 평정 시행의 적절성에 의한 것인지, 아니면 사전에 분할점수를 이용하여 심사본을 분류한 다음 의도적으로 점수를 배분한 결과에 의한 것인지에 대해서는 보다 정밀한 분석이 필요할 것이다. 구체적인 연구 문제에 대한 결론을 제시하면 다음과 같다.

첫째, 각각의 분류일치도 추정 방식에 따른 분류일치도 추정 결과를 살펴보면, 일치도 지수(p)와 Kappa 계수는 모든 책에서 0.6 내외의 양호한 수준을 보였다. 추정 방식 측면에서는 역시 CM 방식의 값이 다소 크게 추정된 점을 일치도 지수 및 Kappa 계수 모두에서 확인하였다.

둘째, 기본 가정의 적합 정도를 알아보는 모형의 적합성을 측정해 본 결과, NM 방식의 기본 가정인 정상성은 대부분의 책에서 적합하다는 결론을 얻었으며, Livingston-Lewis 방식의 기본 가정인 4모수 베타-이항 모형의 적합성을 알아보기 위하여 유효 검사 길이에 의한 실제 자료와 모형의 자료에 대한 점수 분포도를 그려 확인한 결과, 고등학교 수학 본책 및 익힘책 이외의

대부분의 책에서 적합하다는 것을 확인하였다.

한편, 교과서 검정 심사에서 '적격·부적격' 판정이 점수제 판정 방법으로 바뀌면서 처음 시도되는 분류일치도 분석 연구는 다음과 같은 한계를 가지고 있다.

첫째, 연계된 도서의 두 권을 종합한 효과는 분석하지 않고 개별적인 도서의 판정 심사 결과 자료만 가지고 분석을 하였다. 둘째, NM 방식과 Livingston-Lewis 방식 추정에서는 절대적 기준과 관련된 Φ(phi) 수치를 신뢰도 계수로 활용하여야 하며 이 Φ(phi) 계수는 다변량(multivariate) 일반화가능도 이론에 의하여 추정되어야 한다. 검정 심사의 평정에서는 검정기준의 점수 단계가 많고 단계마다의 점수 차도 매우 크면서 단일 문항으로 구성된 점수대도 있어 다변량 일반화가능도 이론 설계에 의한 Φ(phi) 계수를 산출하기가 곤란하여 위의 두 가지 방식 추정에 α계수를 사용하였다.

둘째, 이 연구에서는 국가수준 학업성취도 평가와 같은 피험자들이 응답한 검사 결과에 대한 일반적인 분석이 아닌, 검정 심사본에 대한 전문가의 평가 결과를 대상으로 분석을 처음으로 시행하여, 교과서 심사에 적용한 분류일치도 추정 방법이 검사 상황과 비교하여 문제가 없는지 추가적으로 연구되고 보완되어야 한다는 것이다.

셋째, 이러한 일반적인 피험자 자료가 아닌 교과서 검정 자료를 사용하고, 이를 Φ가 아닌 α 계수를 사용했기 때문에 일치도 지수가 높게 나온 것은 아닌지 향후 세부적인 분석이 요구된다.

2. 제언

우리나라 초·중등학교 학생들이 사용하는 교과서에 대한 검정 심사의 '적격·부적격' 판정에서는 절대평가의 원칙에 따라 미리 합격본 수를 정해놓지 않고 심사를 시행하고 있는데, 검정 심사본이 검정기준에서 요구하는 수준에 도달하였는지의 여부를 판정하기 위한 심사로서 기 제시한 분할점수에 의하여 분류된 심사본들이 제대로 분류가 되었는지를 알아보고자 하였다.

따라서, 이 연구 결과를 바탕으로 추후에 연구를 확대하여 향후에 시행될 초·중·고 검정 심사에서 타당하면서도 신뢰할 수 있는 분할점수를 새로 정하게 된다면, 우리나라에 새롭게 도입된 교과서 점수제 판정 검정에서 '적격·부적격' 심사의 질을 획기적으로 제고하게 되는 계기가 될 것이다.

또한, 학교 현장에서의 평가에도 적용되고 각종 국가 자격증 부여를 위한 인증시험에도 이러한 시도 즉, 과학적 분할점수 설정과 그에 따른 분류일치도 추정이 연계된다면 진정한 준거참조 평가가 국내에 정착되는데 크게 기여를 할 수 있을 것이다.

이 연구 결과를 바탕으로 교과서 검정 심사에서 적용되어야 할 구체적인 제안 사항을 제시하면 첫째, 검정위원은 영역별로 복수로 위촉하여 1인의 주관적 판단에 의한 점수 판정이 되지 않

도록 한다. 둘째, 검정 심사 전에 충분한 시간을 가지고 검정위원들에 대한 점수 판정 방법에 대한 철저한 교육훈련을 실시하여 위원별 능력의 차이가 평정 결과에 영향을 미치지 않도록 한다. 셋째, 심사가 완료된 후 분할점수에 의하여 분류된 심사 결과의 분류일치도를 추정·분석하여 심사본 분류가 제대로 이루어졌는지 검증하고, 심사 결과 환류(Feedback)를 통한 분할점수 조정, 교육훈련 개선 등을 즉시 조치한다.

참 고 문 헌

경제・인문사회연구회(2008). **경제・인문사회연구회 소관연구기관 평가편람**. 경제・인문사회연구회

박정, 김경희, 김수진, 손원숙, 송미영, 조지민(2006). **국가수준 학업성취도평가 기술보고서**. 한국교육과정평가원 연구보고 RRO2006-4.

성태제(1989). 체육계 실기 고사의 합리적 방법과 문제점에 대한 토론. **교육평가연구, 3**(2), 126-130. 한국교육평가학회

채선희, 김명숙, 양명희, 이봉주, 이재기, 최석진, 강문봉, 김경성, 심영택, 이규민, 이재승, 이주섭, 김도남, 김윤희, 김지연, 김혜숙(2003). **2002년 초등학교 3학년 국가수준 기초학력진단평가 연구**. 한국교육과정평가원 연구보고 CRE 2003-1

AERA, APA, NCME (1999). *Standards for Educational and Psychological Testing.* Washington D. C.: American Psychological Association.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy,* version 1.1(CASMA Research Report No. 9). Iowa City, IA: University of Iowa.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedure for estimating decision consistency for single-administration complex assessments*(CASMA Research Research Report No. 7). Iowa City, IA: The University of Iowa.

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method*(ETS Research Report No. 94-39). Princeton, NJ: Educational Testing Service.

Cizek, Gregory J. (2001). *Setting performance standards*. NJ: Lawrence Erlbaum Associates, 15.

Efron B. (1982). *The jackknife, the bootstrap, the other resampling plans.* Philadelphia: SIAM.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

Huynh, H. (1976). On the reliability of decisions in domain referenced testing. *Journal of Educational Measurement, 12,* 253-264

Johnson, R., & Wichern, D. (2002). *Applied Multivariate Statistical Analysis.* Prentice Hall, Upper Saddle River, NJ.

Lee, W. (2001). *Multinomial error model for tests with polytomous items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. Seattle, WA.

Lee, W. (2007). Multinomial and Compound Multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31,* 255-274.

_____(2008a). *Classification Consistency and Accuracy Under the Compound Multinomial Model(CASMA Research Report No. 13).* Iowa City, IA: University of Iowa.

_____(2008b). *manual for MULTI-CLASS: For multinomial and compound multinomial classification consistency.* Iowa City, IA: University of Iowa.

Livingston. S. A., & Lewis. C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement, 32,* 179-197.

Peng. C. J., & Subkoviak. M. J. (1980). A Note on Huynh's Normal Approximation Procedure for Estimating Criterion-Referenced Reliability. *Journal of Educational Measurement, 17,* 359-368.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of the criterion-referenced test. *Journal of Educational Measurement, 13,* 265-276.

Wan, L. (2008). *Estimating Classification Consistency for Single Administration Complex Assessments Using Non-IRT Procedure.* Unpublished Doctoral dissertation. University of Iowa.

· 논문접수 : 2012-01-01/ 수정본 접수 : 2012-02-07/ 게재승인 : 2012-02-22

ABSTRACT

A Study on Estimating Classification Consistency Using Non-IRT Procedures for Textbook Authorization

Chang-Hwan Kim
(Korea Institute for Curriculum and Evaluation)

A testee's pass or fail in test is determined by the same standard(cut-off score) that requires over arbitrary mean scores(80, 70 or 60 points etc.) in overall area without any systematic standard-setting procedure in Korea. There are a few procedures estimating classification consistency with tests consisting of only dichotomous items by using Item Response Theory(IRT) procedure. Undoubtedly, there has not been any case of setting standard and estimation classification consistency in "Textbook Authorization". In these poor circumstances, this study aimed to investigate the performance of three procedures for estimating classification consistency for complex assessments and data from the "Textbook Authorization Screening of Mathematics and English" were used. The summary of the major conclusions identified by this study is as follows.

First, the three procedures that this study employed in five non-IRT estimating classifying consistency are a normal approximation procedure(NM), the Livingston-Lewis procedure(LL) and a compound multinomial procedure(CM). Estimates of agreement index p and Kappa yielded by every procedure were somewhat high, especially, in CM procedure and Mathematics textbook for high school students. Second, normality is an important assumption for NM procedure. To investigate whether this assumption holds for the data, Q-Q plot was constructed. As a result, the points lay very near along a straight line, indicating that marginal normality assumption was tenable. The four-parameter beta-binomial model is an integral part of the LL procedure. To investigate the fit of the model for the data, observed and fitted distribution of effective test length scores are plotted. Due to the small sample size, there were many irregularities in the observed distributions, especially in Mathematics textbook for high school students, yet, roughly the observed and fitted distributions shared the same shape. Finally, several proposals for the textbook authorization were summarized, as well as limitations of the study and future research possibilities.

Key Words : Criterion-referenced test, Classification consistency, Agreement index p, Kappa, Normal-Approximation procedure, Livingston-Lewis procedure, Compound multinomial procedure, Textbook authorization