교육과정평가연구
The Journal of Curriculum and Evaluation
2014, Vol. 17, No. 3, pp. 223~245

수준설정의 타당성 검증을 위한 평정자 기대 정답률의 예측오차 분석

박 인 용(한국교육과정평가원 부연구위원)*
송 미 영(한국교육과정평가원 연구위원)**
김 성 숙(한국교육과정평가원 선임연구위원)

≪ 요 약 ≫

이 연구는 앙고프(Angoff) 방법에 의한 수준설정의 절차적 측면에서 타당성을 점검하는 연구로, 평정자의 기대 정답률 예측에 대한 라운드별 예측 정확성을 탐색하였다. 2010년 중·고등학교 학업성취도 평가의 수준설정에서 라운드별 평정자가 각 문항별로 예측한 기대 정답률 자료와 문항 응답 자료를 활용하여 기대 정답률 예측의 정확성 지수를 문항 난이도, 문항 유형별로 산출한 결과, 수준설정 라운드가 진행됨에 따라 평정자들의 기대 정답률과 실제 정답률간 오차가 점차 감소하여 절차적 측면에서의 타당성 증거를 확인하였다. 평정자들은 문항이 쉬울 경우 기대 정답률을 과소 추정하였고, 문항이 어려울 경우 과대 추정하는 경향을 보였으며, 문항 유형에 따라서는 학교급, 교과, 성취수준별로 라운드에 따른 오차 경향이 다양하게 산출되었다. 이러한 결과에 기반하여 내용 전문가의 문항 정답률 예측을 필요로 하는 수준설정 방법을 적용할 때 기대 정답률 예측에 대한 평정자 훈련의 고려사항을 제언하였다.

주제어: 수준설정, 앙고프 방법, 기대 정답률, 절차적 타당성

* 제1저자, iypark@kice.re.kr
** 교신저자, mysong@kice.re.kr

Ⅰ. 연구 배경 및 목적

최근의 교육평가 분야에서 학교 및 학생의 수행 비교와 서열에 중점을 두기보다는 특정 준거의 달성 정도를 평가하는 준거참조평가(criterion referenced assessment)에 대한 관심은 국가수준뿐 아니라 학교수준에서 지속적으로 증가하고 있다. 국가수준 학업성취도 평가(박정 외 2006; 이하 '학업성취도 평가')나 초등학교 3학년 국가 수준 기초 학력 진단 평가(이인제 외 2004)에는 평가의 목적에 부합되도록 평가 계획 당시부터 준거참조평가가 적용되어 왔으며, 최근 대학수학능력시험의 일부 영역에 준거참조평가를 도입하는 방안에 대한 논의가 활발히 진행되고 있다(강태중, 2014; 김신영, 2009; 교육부, 2013; 한국교육개발원, 2014). 준거참조평가의 기능은 국가수준의 대규모 평가에서뿐만 아니라 학교 단위의 평가에서도 강조되고 있다. 학교수준에서는 성취평가제가 2012년에 중학교 1학년부터 본격적으로 도입되고, 고등학교에서 시범 운영되는 등 교사에 의한 학생 평가에 준거참조평가가 적용되고 있다.

준거참조평가는 특정 준거에 어느 정도 도달했는지를 평가하는 것으로 평가 결과는 도달 정도에 따라 도달/미도달 또는 우수학력/보통학력/기초학력/기초학력 미달 또는 A/B/C/D/E 등의 성취수준으로 개별 학생들과 교사들에게 제공된다. 이와 같은 방식으로 결과를 보고하는 평가 체제에서는 학생들의 성취수준을 구분하기 위한 '분할점수(cut score)'가 필요하다. 분할점수는 어떤 척도 위에 설정되는 특정 점수를 지칭하고, 분할점수를 기준으로 구분되는 성취수준은 피험자가 알고 있거나 할 수 있는 수행능력의 준거가 된다(Kane, 2001). 또한, 검사 점수를 몇 개의 구간으로 나누기 위해 어떤 척도 위에 한 개 이상의 분할점수를 설정하고 각 구간에 속한 피험자의 수행수준 특성을 작성하는 일련의 작업을 '수준설정(standard setting)'이라 하며, 수행수준 또는 성취수준은 학생들이 실제로 알고 있거나 할 수 있는 것에 대한 정의와 구체적인 사례를 제시한 것을 의미한다(한국교육평가학회, 2004). 예를 들어, 국가수준 학업성취도 평가의 경우, 점수 척도 위에 세 개의 분할점수가 설정되어 점수 척도는 네 개의 구간으로 나누어진다. 점수 척도 위에 구분된 네 구간은 각각 우수학력 수준, 보통학력 수준, 기초학력 수준과 이에 도달하지 못한 기초학력 미달을 나타낸다.

수준설정 방법은 많은 학자들에 의해 다양한 방법들이 개발되고 사용되어 왔으며, 앙고프 방법(Angoff methods; Angoff, 1971) 및 변형된 앙고프 방법(modified Angoff methods; Zeiky, 2001), 북마크 방법(bookmark method; Lewis, Mitzel, Green, & Patz, 1999)이 가장 널리 사용되고 있다. 이와 같은 다양한 수준설정 방법들 중에서 분할점수를 정확하게 선정하기 위한 간단하고 분명한 방법은 없으며(Jaeger, 1992; Kane, 2001), 대부분의 방법에서 검사 내용, 문항, 학생에 대한 전문가인 평정자들의 특정 성취수준별 학생에 대한 기대 정답률을 예측하는 것이 요구되고 예측 과정에 주관성이 개입될 가능성이 다수 존재한다(Glass,

1978; Shepard, 1979). 따라서 수준설정에 있어서 가장 중요한 것은 '명확한 의미를 갖는' 분할점수를 산출하는 것이며, 가능한 한 전문가에게나 일반 대중에게 납득될 수 있어야 한다. 이와 같은 명확한 의미의 분할점수를 기반으로 검사 목적 등 다양한 상황에 따라 적절한 수준설정 방법이 사용될 수 있다. 이러한 관점에서 특정한 방법을 사용하여 성취수준을 구분하기 위한 분할점수를 설정한 이후 분할점수의 타당성을 확인하는 작업이 필수적으로 이루어져야 한다(김석우 외, 1999; 이규민, 2004).

여러 가지 수준설정 방법 중에서 '변형된 앙고프 방법'은 비교적 간단하고 절차로 선다형뿐 아니라 서답형 및 혼합형 검사에 적용될 수 있는 장점이 있어 대규모 평가인 학업성취도 평가 등에서 활용할 뿐 아니라(박정 외, 2006; 김성숙 외, 2010; Allen et al, 2001) 성취평가제를 위해 학교에서 적용할 수 있는 수준설정 방법 중 하나로 교사들에게 권고되고 있다(한국교육과정평가원, 2012, 2014). '변형된 앙고프 방법'은 평정자들이 성취수준별 최소능력자(minimally acceptable person)를 개념화하고 최소능력자 집단이 각각의 문항에 대해 정답할 확률을 예측한다. 이는 일종의 기대 정답률로 각 문항에 대해 예측된 기대 정답률의 합으로 최종 분할점수를 설정한다(Cizek & Bunch, 2007). 이러한 측면에서 평정자들의 기대 정답률에 대한 예측 및 이에 대한 논의와 합의는 분할점수 설정에 매우 핵심적인 역할을 한다. 수준설정 과정에서 반복적인 논의를 통해 평정자들 간 합의된 최소능력자의 특성은 평정자들의 기대 정답률에 반영되며, 이는 분할점수로 분류된 실제 학생들의 특성과 유사해야 한다. 그러나 평정자들의 기대 정답률 예측은 다양한 요인에 의해 영향을 받을 수 있으며(Impara & Plake, 1998), 문항 유형이나 문항 난이도 등의 영향으로 인해 수준설정에서의 최소능력자에 대한 특성과 실제 분류되는 학생들의 특성에 차이가 나타날 수 있다. 이러한 차이는 수준설정의 타당성을 결여시키는 요인이 될 수 있다. 따라서 수준설정 과정에서 논의된 최소능력자의 특성과 실제 분할점수대의 점수를 획득한 학생들의 수행 특성의 차이를 확인해야 하며, 이를 위해 평정자들의 기대 정답률과 실제 분할점수를 획득한 학생의 정답률의 차이를 점검해야 할 필요가 있다.

본 연구는 수준설정, 특히 변형된 앙고프 방법을 적용한 수준설정에서의 타당성을 확인하기 위해 분할점수 산출에 주요하게 영향을 미치는 평정자의 기대 정답률 예측에 대한 정확성과 문항의 특성에 따른 기대 정답률의 예측 오차를 분석하였다. 평정 라운드별 기대 정답률의 오차 변화를 점검함으로써 수준설정 과정인 라운드에 따른 평정자 간 논의 및 산출된 성취수준의 타당성을 점검하는 것을 목적으로 하고 있다. 이를 기반으로 수준설정에 있어 평정자들의 기대 정답률 예측에 대한 훈련 및 정확성 제고를 위한 시사점을 제공하고자 하였다.

Ⅱ. 수준설정 타당화

성취수준을 구분하기 위한 분할점수를 설정하는 방법에는 수십 가지가 넘는 다양한 방법들이 있으며, 이러한 방법들은 크게 검사도구 내용 분석 평가에 의한 방법, 피험자 특성 평가에 의한 방법 등으로 구분된다(Cizek, 1996; Kane, 1994). 검사도구 내용 분석 평가에 의한 방법은 평정자들이 검사의 각 문항을 검토하면서 특정 수준의 경계선에 있는 학생들의 문항에 대한 기대 수준에 초점을 두어 분할점수를 설정한다. 이 방법에서 수준을 설정하는 전문가, 즉 평정자들은 특정 수준의 경계선 혹은 수준 내 최소능력을 가진 학생의 수행 능력을 개념화하고, 검사 내 각각의 문항에 대하여 내용을 분석 및 평가함으로써 수준을 설정한다. 이와 같은 방법 중 앙고프 방법, 북마크 방법이 가장 널리 사용되고 있다.

앙고프 방법은 전문가들이 검사도구의 내용 분석을 통해 분할점수를 설정하며, 가장 오래되고 안정적인 방법이다. 이 방법은, 가설적으로 상정한 최소능력자들이 각 문항에 정답할 확률을 추정하고 문항별 비율의 총합으로 분할점수를 산출한다. 앙고프 방법에서는 평정자의 수나 자격, 평정자들이 문항의 정답 확률을 예측할 때 수정할 수 있도록 허용할 것인지, 또는 문항의 정답이나 채점 기준 등을 제공할 것인지 등과 같은 분할점수 설정 절차나 과정에 대한 구체적인 내용에 대하여는 언급되지 않는다. 따라서 수준설정 과정에서 판단의 반복 과정, 경험적인 데이터의 제공 등 구체적인 절차 등이 변형되어 지속적으로 활용되어 왔다(Zieky, 2001). 이러한 변형된 앙고프 방법은 절차가 비교적 간단하고 문항을 맞힐 확률을 다양하게 예측할 수 있으며 선다형 문항뿐 아니라 서답형 문항, 혼합형 검사에도 적용될 수 있는 장점이 있다. 그러나 앙고프 방법과 같은 최소능력자의 개념에 대한 의존성이 높은 수준설정 방법은 분할점수 산출에 있어 평정자 간의 최소능력자에 대한 개념의 공유와 평정자들의 기대 정답률 예측의 정확성 제고가 매우 중요하게 작용한다. 이러한 이유로 Kane(1986, 1994)와 Impara와 Plake(1998)은 성취수준의 타당성을 확보하기 위해, 즉, 평정자들이 개념화한 최소능력자의 특성이 성취수준에 정확하게 반영되었는지를 확인하기 위해 실제 성취수준 간 경계선 집단 학생들의 문항에 대한 수행수준과 비교하여 점검해야 할 것을 제안하고 있다. 앙고프 방법뿐 아니라 문항 혹은 학생에 대한 평정자들의 분석과 판단을 활용하는 수준설정 방법들은 평정자들의 주관성, 절차 상의 다양성 등이 존재하며 수준설정을 평가하기 위한 절대적인 준거가 없는 상황에서 수준설정에 대한 타당성 점검은 필수적이다.

Kane(1994, 2001)은 수준설정에 있어 절차적(procedural), 내적(internal), 외적(external) 측면에서 타당성을 검증할 것을 제안하고 있다. 절차적 측면의 타당성은 수준을 설정하는 절차가 적절하였는지에 대한 것으로, 수준설정 참여자와 그 과정을 살펴봄으로써 확인할 수 있다. 절차적 타당성은 수준설정이 실제 이루어지기 이전에 수준설정 과정이 명확하고 분명하게 정의

되어 있는 정도를 기준으로 확인하거나(van der Linden, 1995), 수준설정 과정 및 자료 분석 과정의 정확성과 분석 결과의 해석에 대한 정확성을 확인함으로써(Berk, 1986) 점검할 수 있다. 또한, 평정자들의 훈련 과정, 최소능력자의 개념화에 대한 명확성, 수행수준 진술의 명확성과 평정자들이 수준설정에 대한 편의성 측면에서의 피드백 정도(Kane, 1994; 2001), 평정자의 수행수준 예측의 정확도(Impara & Plake, 1998)를 준거로 수준설정에 대한 절차적 타당성을 확인할 수 있다.

수준설정의 내적인 측면에 대한 타당성은 분할점수 산출에 있어서 방법 간, 평정자 간 혹은 평정자 내 일관성과 관련된다. 이를 점검하기 위해 Cizek(1996), Kane(1994, 2001), van der Linden(1995)은 동일한 수준설정 방법이 반복되어 적용되었을 때 수행수준 추정의 정확성 정도와 피험자들이 동일한 성취수준으로 분류되는 정도, 평정자들의 반복된 평정에 대한 일관성 정도를 확인하였고, Kane(1995), Shepard 외(1993)은 내용영역, 인지영역 등에서의 수행수준에 대한 일관성 정도를 분석하였다. 또한, 김성숙 외(2012)에서는 다변량 일반화가능도 이론의 접근을 통해 수준설정 과정에서 평정자, 문항, 라운드 오차요인의 상대적 영향력을 탐색하고, 기대 정답률 예측의 신뢰도를 확인하였고, 강태훈 외(2011)에서는 검사 길이 및 수준의 개수에 따른 성취수준 분류의 일관성을 점검하였다. 이현숙(2007)과 송미영 외(2013)는 다양한 방법을 적용하여 분할점수의 분류 일관성과 정확성 지수를 통해 성취수준 분류에 대한 신뢰도를 확인함으로써 Kane(1994, 2001)에서 제시하고 있는 수준설정에 대한 내적인 측면에서의 타당성을 점검하였다.

외적인 측면의 타당성은 다른 종류의 객관적 정보와의 비교 또는 다른 수준설정 방법과의 비교에서 일치하는 정도를 통해 수준설정의 타당성을 확인하는 것이다. 양길석(2000), Jaeger(1992)와 Kane(1994, 2001), Lee 외(2010)에서는 서로 다른 방법을 적용하여 산출된 수준설정 결과가 일치하는 정도를 확인함으로써 외적 타당성을 점검하였고, Berk(1996), Giraud 외(2000), Kane(1994, 2001), Shepard 외(1993)와 민경석(2013)은 수준설정의 외적 타당성을 실제 적용된 검사와 유사한 다른 검사와의 분할점수나 성취수준 비율 등의 결과 비교를 통해 점검하였다. 또한, 박연복, 이규민과 강상진 (2011)은 북마크 방법을 통해 산출된 분할점수와 군집분석을 적용한 수준설정에서의 분할점수를 비교하였고, 백순근과 최인희(2006)은 Rasch 방법과 앙고프 방법 간 분할점수를 비교하였으며, 장윤선과 성태제(2009)는 문항반응이론에 기초한 수준설정 간 방법 비교를 통하여 외적 타당성을 점검하고자 하였다.

이 연구의 목적은 변형된 앙고프 방법을 적용한 수준설정의 절차적 측면에 대한 타당성 점검함에 있어 실제 분할점수를 받은 학생들의 문항 정답률과 평정자들이 수준설정 과정에서 예측한 기대 정답률을 비교 분석함으로써 평정자 간 논의의 타당성과 함께 기대 정답률 예측의 정확성을 확인하고자 하였다.

Ⅲ. 연구 방법

1. 분석대상

변형된 앙고프 방법을 통한 수준설정에서 평정자들의 기대 정답률 예측의 정확성을 분석하기 위해 국가수준의 대규모 평가인 학업성취도 평가의 수준설정 자료를 활용하였다. 2010년 중학교 국어, 수학, 영어 교과에 대한 수준설정 과정의 평정자 기대 정답률 예측 자료와 학생 응답의 채점 자료를 활용하여 실제 문항 정답률을 산출하고, 이를 기준으로 수준설정에서 평정자들의 기대 정답률에 대한 예측 정확성을 검증하였다.

먼저, 학업성취도 평가 수준설정에서의 평정자와 기대 정답률 예측 과정을 살펴보면, Cizek (1996), Raymond와 Reid(2001) 등의 기준에 따라 학업성취도 평가의 대상 학년, 내용 영역에 대한 전문가 20명으로 구성되었다. 교과별로 정의된 성취기준을 토대로, 평정자들이 성취수준별 최소능력자를 개념화하고, 문항, 정답 및 채점 기준을 면밀히 검토한 후 평정의 편이성을 위해 5% 단위로 정답률 예측이 이루어졌다. 이 때, 평정자들은 기초학력, 보통학력, 우수학력의 최소능력자들이 문항에 정답할 확률을 각자 독립적으로 예측하고, 결과를 평정자들 간 비교, 논의하였으며, 이러한 과정을 국어와 수학의 수준설정에서는 3라운드에 걸쳐, 영어 교과의 경우는 4라운드에 걸쳐 모든 문항에 대해 반복적으로 수행하였다. 1라운드에서는 최소능력자의 개념에 대한 논의만 진행된 상태로 문항에 대한 기대 정답률 예측이 이루어졌으나, 2라운드부터는 학생 전체 집단의 실제 정답률과 답지반응분포 등 학생들의 실제 수행수준에 대한 정보를 제공하였다.

이 연구에서는 라운드별 기대 정답률의 오차 변화를 성취수준 간 구분선(또는 경계선)인 분할점수를 획득한 학생 집단(이하 분할점수 집단)과 성취수준에 포함되는 모든 학생 집단에서의 실제 정답률을 기준으로 산출하였다. 또한, 학생 전체 집단의 정답률에 의한 문항 난이도에 따라 라운드별 기대 정답률의 오차 변화와 문항 유형별 기대 정답률 오차 변화를 분할점수 집단에서의 실제 정답률을 기준으로 탐색하였다. 교과별 분할점수 집단과 전체 학생 수는 〈표 1〉과 같다.

〈표 1〉 각 교과의 성취수준별 학생 수

성취수준		국어	수학	영어
분할점수 집단	우수학력	10,383	11,368	12,814
	보통학력	8,141	8,375	8,300
	기초학력	2,321	1,637	1,163
전체 집단		658,040	658,279	658,245

문항 난이도 수준은 문항의 정답률 예측에 영향을 주는 요인이 될 수 있으므로 이 연구에서는 문항별 전체 집단의 정답률의 분포를 고려하여 정답률이 0.4 미만일 경우 '난이도 상', 0.4 이상 0.7 미만일 경우 '난이도 중', 0.7 이상일 경우 '난이도 하'로 문항을 구분하여 분석하였다. '난이도 상'은 전체 학생의 40% 미만이 답을 맞히지 못한 어려운 문항에 해당하며, '난이도 하'는 전체 학생의 70% 이상이 정답한 쉬운 문항에 해당된다. 이 연구에서는 이와 같이 문항 난이도를 세 수준으로 구분하여 각각의 난이도 수준에서 기대 정답률 예측의 정확성 정도가 어떠한지를 확인하였다.

학업성취도 평가의 각 교과별 검사 구성은 선다형-이분 문항과 서답형-다분 문항이 혼합되어 있다. 각 학교급 및 교과별 검사의 문항 난이도 및 유형별 문항 수는 〈표 2〉와 같다.

〈표 2〉 중학교 교과별 검사의 문항 난이도 및 유형별 문항 수

교과	문항유형	난이도 상	난이도 중	난이도 하	소계
국어	선다형	8	14	6	28
	서답형	1	3	2	6
	소계	9	17	8	34
수학	선다형	4	17	8	29
	서답형	5	3	0	8
	소계	9	20	8	37
영어	선다형	5	23	6	34
	서답형	4	2	0	6
	소계	9	25	6	40

교과별 전체 문항 수는 다소 차이가 있지만 34~40개의 문항으로 이루어져 있고, 선다형 문항은 28~34개, 서답형 문항은 하위 문항 단위로 6~10개로 전체 문항 수의 약 20~30% 정도이다. 서답형 문항의 경우 하위 문항이 포함되어 있는데, 2010년 수준설정 과정에서 문항에 대한 기대 정답률 예측에 있어 하위 문항별로 정답률을 예측하여 실제 정답률 산출에도 하위 문항 단위로 정답률을 산출하였다. 또한, 선다형 및 서답형 등 문항 유형에 따른 기대 정답률 오차의 경우 서답형 문항의 난이도 분포가 상대적으로 높게 나타나, 문항 난이도에 따른 영향을 통제하기 위해 '난이도 하'인 문항을 제외한 후 문항 유형별 예측 정확성을 탐색하였다.

2. 분석방법

이 연구에서는 수준설정에서 문항의 난이도, 문항 유형에 따른 평정자들의 기대 정답률 예측의 정확성을 분석하기 위해 평정자들이 예측한 기대 정답률과 분할점수 집단의 실제 정답률과의

차이를 통해 오차를 산출하였다. 분할점수 집단은 각 성취수준을 구분하는 분할점수를 획득한, 즉 성취수준 간 경계선 상의 학생 집단이다. 성취수준 분할점수와 분할점수 집단은 평정자들이 수준설정 과정에서 최소능력자 집단으로 개념화한 것이 기대 정답률을 통해 실제로 발현된 것이며, 이 연구에서는 분할점수 집단의 정답률과 기대 정답률의 차이를 예측 오차로 제시하였다. 즉, 이 연구에서의 예측 오차는 최소능력자 집단의 정답률 추정치인 기대 정답률과 실제 정답률의 차이를 의미한다. 이를 위해 다음과 같은 세 가지 지표를 사용하였다.

$$SE_{ir} = \sqrt{\frac{1}{n_j}\sum_{j=1}^{n_j}(\widehat{P_{irj}} - P_i)^2} \quad , \qquad SE_r = \frac{1}{n_i}\sum_{i=1}^{n_i} SE_{ir}$$

$$BIAS_{ir} = \sqrt{\frac{1}{n_j}\sum_{j=1}^{n_j}(\widehat{\hat{P}_{ir}} - P_i)^2} \quad , \qquad BIAS_r = \frac{1}{n_i}\sum_{i=1}^{n_i}|BIAS_{ir}|$$

$$SEE_{ir} = \sqrt{\frac{1}{n_j}\sum_{j=1}^{n_j}(\widehat{P_{irj}} - \widehat{\overline{P_{ir}}})^2} \quad , \qquad SEE_r = \frac{1}{n_i}\sum_{i=1}^{n_i} SEE_{ir}$$

위 식에서 i는 문항을 나타내며, r은 수준설정에서의 라운드, j는 평정자를 나타낸다. P_i는 특정 분할점수를 획득한 학생 집단에서 i문항의 실제 정답률을 나타내며, $\hat{P}_{irj}$는 i문항에 대해 r번째 라운드에서 평정자 j가 최소능력자들의 정답률을 예측한 기대 정답률을 나타낸다. $RMSE_{ir}$은 평정자들이 i문항에 대해 r번째 라운드에서 예측한 기대 정답률과 실제 정답률의 전반적인 오차를 보여준다. $BIAS_{ir}$은 i문항에 대해 r번째 라운드에서 예측한 평정자들의 기대 정답률 평균과 실제 정답률 간 차이로 평정자들의 평균적인 예측오차를 나타낸다. SEE_{ir}은 평정자들이 i문항에 대해 r번째 라운드에서 예측한 기대 정답률의 표준편차로 각 문항에 대해 라운드별 평정자 간 기대 정답률에 대한 논의의 수렴 정도를 확인할 수 있다.

기대 정답률과 실제 정답률의 전반적인 오차를 나타내는 $RMSE_{ir}$는 평정자들의 평균 예측 정확성과 평정자 간 의견 수렴 정도를 모두 반영하고 있으며, 평정자들의 기대 정답률 평균과 실제 정답률과의 차이, 평정자들 간 기대 정답률의 차이로 분할할 수 있다. 즉, $RMSE_{ir}^2 = BIAS_{ir}^2 + SEE_{ir}^2$의 관계를 가진다. $RMSE_{ir}$, $BIAS_{ir}$, SEE_{ir}는 라운드별로 모든 문항에 대해 산출하였으며, 이로써 라운드에 따른 각 문항별 기대 정답률의 오차 변화와 평정자 간 기대 정답률 수렴 정도를 확인하였다. 또한, 각 라운드별 문항에 대한 $RMSE_{ir}$, $BIAS_{ir}$, SEE_{ir}의 평균을 통해 각 라운드에서 평균적인 기대 정답률의 오차 정도와 평정자 간 기대 정답률 수렴 정도를 확인할 수 있다. 이 연구에서는 교과별・라운드별・문항 난이도별・문항 유형별로 각 성취수준에 대한 세 가지 지표, $RMSE_{ir}$, $BIAS_{ir}$, SEE_{ir}를 산출하여 각 성취수준별 분할점수를 획득한 집단의 정답률을 기준으로 하여 성취수준에서 평정자의 기대 정답률 예측에 대한 정

확성을 탐색하였으며, 문항 난이도 수준(상, 중, 하) 및 문항 유형(선다형 , 서답형)에 따라 기대 정답률 예측의 정확성을 각각 비교 분석하였다.

Ⅳ. 연구 결과

1. 평정 라운드에 따른 기대 정답률 예측 오차

2010년 중학교 학업성취도 평가 분할점수 집단과 성취수준별 전체 집단의 문항 응답을 통해 산출한 문항별 실제 정답률을 기준으로 산출한 수준설정 과정에서 라운드별 평정자의 기대 정답률의 RMSE, BIAS, SEE는 〈표 3〉과 같다.

〈표 3〉 교과별 성취수준 라운드에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

교과	성취수준	준거	분할점수 집단				성취수준별 전체 집단			
			1R	2R	3R	4R	1R	2R	3R	4R
국어	우수	RMSE	19.38	9.57	9.37	-	16.18	10.57	10.50	-
		BIAS	14.68	7.88	7.79	-	10.74	9.38	9.40	-
		SEE	10.65	4.00	3.82	-	10.65	4.00	3.82	-
	보통	RMSE	22.97	11.19	11.40	-	22.42	13.29	15.16	-
		BIAS	17.34	9.28	9.77	-	17.52	11.36	13.74	-
		SEE	12.76	5.20	4.68	-	12.76	5.20	4.68	-
	기초	RMSE	12.33	7.17	7.31	-	19.99	17.66	16.72	-
		BIAS	6.89	4.85	5.38	-	15.98	16.49	15.57	-
		SEE	9.58	4.75	4.23	-	9.58	4.75	4.23	-
수학	우수	RMSE	12.15	10.60	10.28	-	9.79	11.65	11.64	-
		BIAS	9.14	8.27	8.43	-	6.77	9.88	10.32	-
		SEE	6.41	5.62	4.87	-	6.41	5.62	4.87	-
	보통	RMSE	18.05	15.20	14.58	-	20.90	20.78	20.82	-
		BIAS	13.23	12.07	12.21	-	16.89	18.49	19.09	-
		SEE	10.66	7.87	6.54	-	10.66	7.87	6.54	-
	기초	RMSE	8.75	8.15	7.82	-	19.46	18.51	17.94	-
		BIAS	5.98	6.01	5.97	-	18.01	17.43	17.03	-
		SEE	5.42	4.61	4.14	-	5.42	4.61	4.14	-

교과	성취 수준	준거	분할점수 집단				성취수준별 전체 집단			
			1R	2R	3R	4R	1R	2R	3R	4R
영어	우수	RMSE	15.69	13.84	13.06	12.72	9.93	16.35	14.91	14.50
		BIAS	13.53	11.24	11.58	11.54	7.17	14.31	13.71	13.57
		SEE	6.21	7.09	5.25	4.61	6.21	7.09	5.25	4.61
	보통	RMSE	23.75	16.31	12.86	9.96	17.89	18.57	17.48	19.11
		BIAS	19.04	12.06	10.02	7.99	11.32	14.52	15.11	18.30
		SEE	12.25	10.09	7.23	4.89	12.25	10.09	7.23	4.89
	기초	RMSE	14.92	10.25	9.68	9.47	15.82	15.05	14.20	13.96
		BIAS	10.54	7.94	8.23	8.24	10.99	13.82	13.36	13.22
		SEE	9.84	5.04	3.81	3.61	9.84	5.04	3.81	3.61

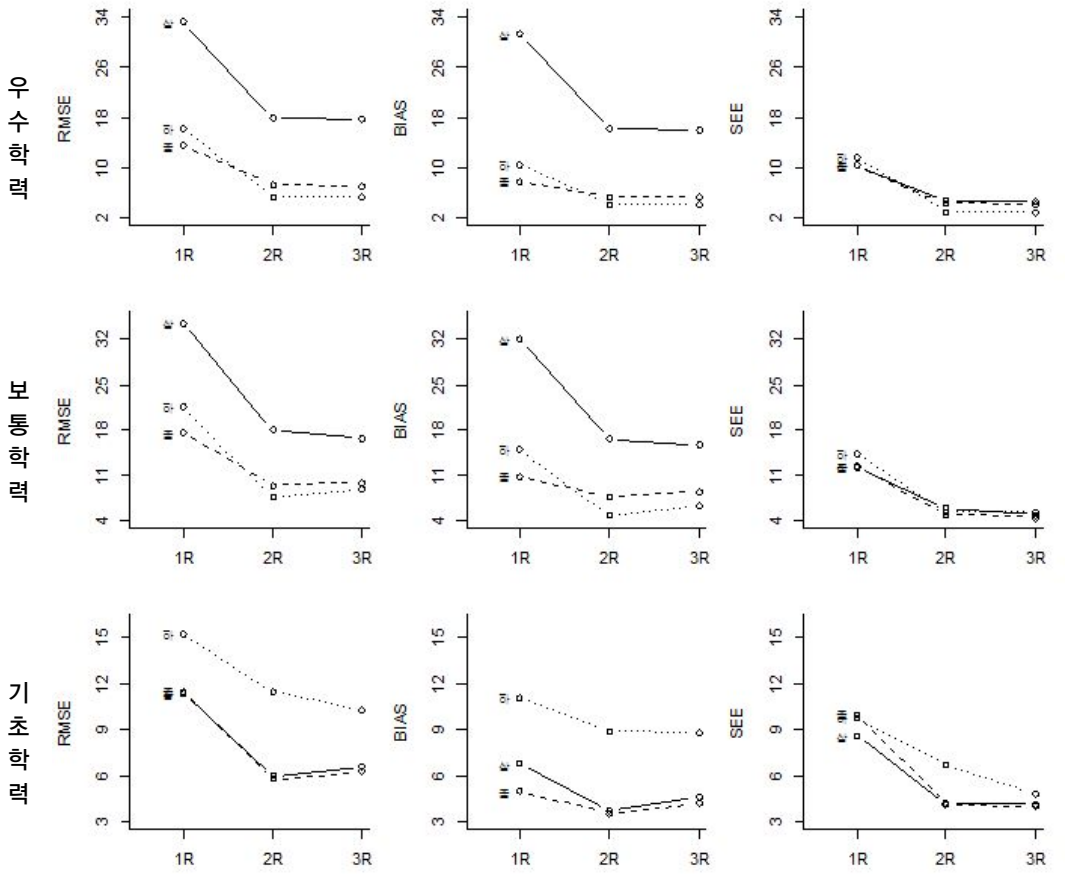
교과별 라운드에 따른 RMSE의 결과를 보면, 모든 교과에서 라운드에 따라 감소하고 있었으며 그 정도와 경향은 교과별, 성취수준별로 차이가 있었다. 보통학력에서 RMSE가 가장 크게 나타났으며, 우수학력, 기초학력 순으로 나타나 보통학력에서 기대 정답률 예측의 전반적인 오차정도가 우수학력과 기초학력에 비해 상대적으로 크고, 기초학력의 수준설정에서 가장 오차가 작은 것을 알 수 있다.

평정자들의 평균적인 예측 오차를 보여주는 BIAS를 보면, RMSE의 경향과 유사하게 나타났는데, 국어와 영어의 경우 라운드가 진행될수록 오차가 감소하였으나, 수학의 경우에는 라운드에 따른 오차의 변화는 크지 않았다. SEE는 모든 교과에서 라운드가 진행될수록 감소하여 라운드에 따라 평정자들의 기대 정답률에 대한 수렴 정도가 높은 것을 알 수 있다. 특히, 1라운드에서 2라운드로 넘어가는 시점에 오차가 가장 많이 줄어들었는데 이러한 결과는 표집 학생들의 실제 수행수준에 대한 정보를 제공함에 따라 2라운드에서 성취수준별 최소능력자의 개념에 대한 평정자 간 차이가 크게 줄어들었고, 라운드를 거듭하여 논의가 진행될수록 평정자 간 의견이 점차 수렴되었음을 보여준다. 또한, 1라운드 결과를 살펴보면, 분할점수 집단 정답률을 기준으로 했을 때의 결과가 성취수준 집단의 정답률을 기준으로 했을 때보다 상대적으로 크게 나타난 것을 볼 수 있다. 이러한 점은 평정자들은 성취수준별 최소능력자의 개념화에 있어 성취수준 전체 구간의 평균쪽으로 개념화한 것을 간접적으로 보여준다.

2. 문항 난이도에 따른 기대 정답률 예측 오차

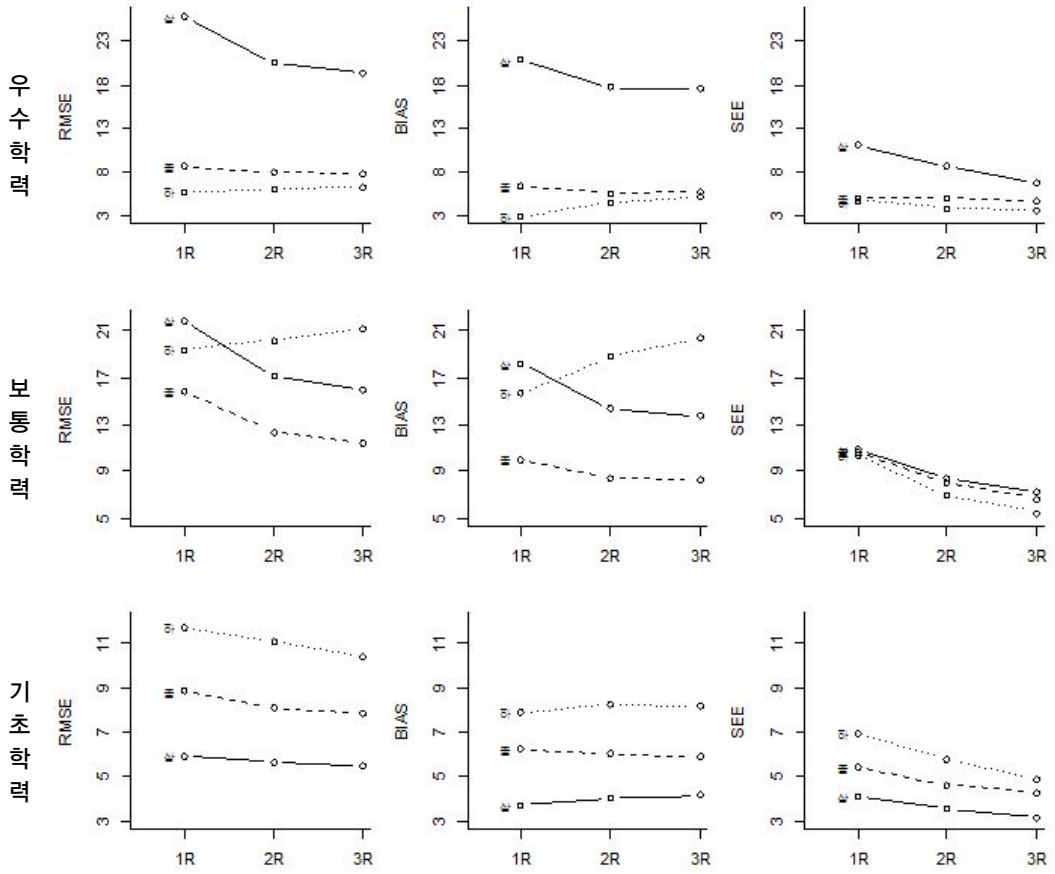
수준설정 과정에서 기대 정답률 오차의 라운드별 변화가 문항 난이도에 따라 차이가 나타나는지 확인하기 위해 전체 집단의 문항 정답률을 통해 문항의 난이도를 상, 중, 하로 구분하고,

기대 정답률 오차의 변화를 탐색하였다. 교과별 문항 난이도에 따른 기대 정답률의 RMSE, BIAS, SEE는 〔그림 1〕~〔그림 3〕과 같다.



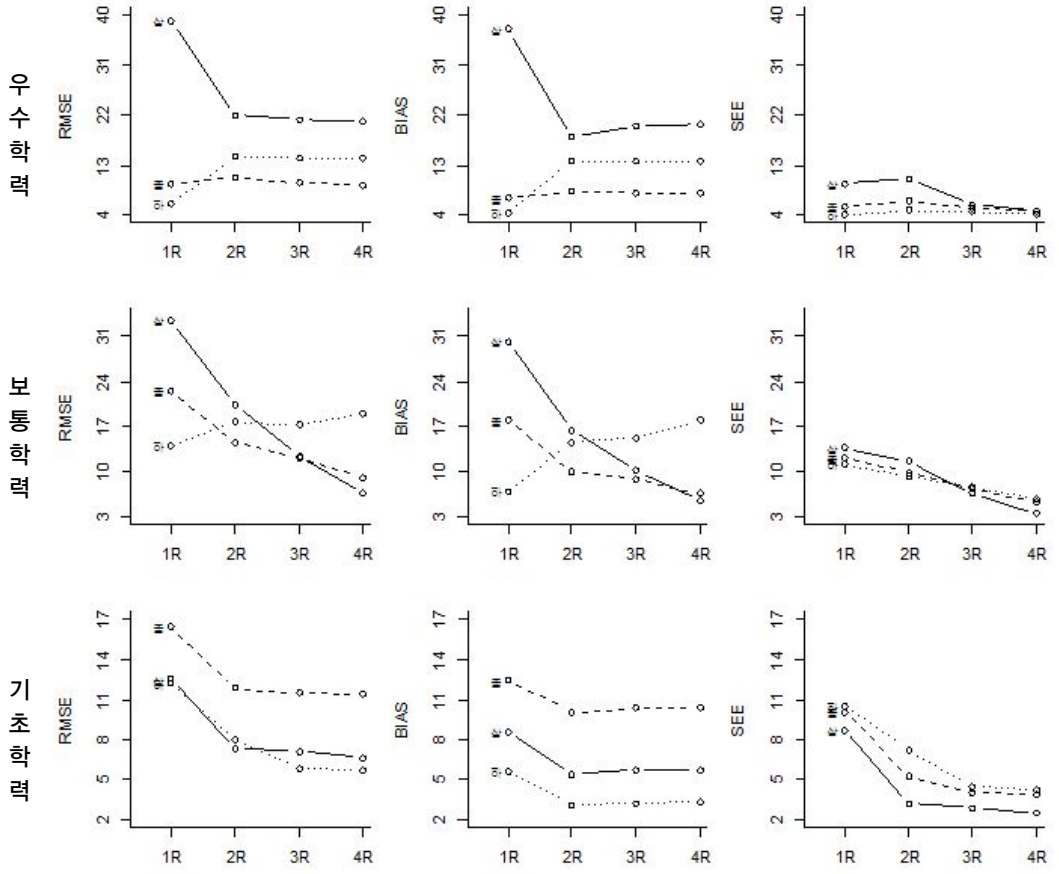
〔그림 1〕 국어 성취수준별 문항 난이도에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

문항의 난이도를 상, 중, 하의 세 수준으로 구분하고, 분할점수 집단의 정답률을 기준으로 기대 정답률의 오차를 나타내는 RMSE, BIAS와 평정자들 간의 의견 수렴 정도를 보여주는 SEE를 교과별로 살펴보면, 국어의 경우 우수학력과 보통학력 설정에서는 어려운 문항, 기초학력 설정에서는 쉬운 문항에서 오차가 상대적으로 크게 나타났다. SEE의 경우 문항의 난이도에 따른 경향의 차이는 크지 않았으며, 전반적으로 모두 라운드가 진행될수록 SEE가 줄어들어 평정자 간 의견 수렴 정도가 높게 나타난 것을 알 수 있다.



〔그림 2〕 수학 성취수준별 문항 난이도에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

수학의 우수학력 설정에서는 어려운 문항에 대한 기대 정답률 예측 오차가 상대적으로 매우 크게 나타났으며, 평정자 간 의견 수렴 정도도 난이도 중과 난이도 하의 문항에 비해 상대적으로 작게 나타났다. 보통학력 설정에서는 난이도 상과 중인 문항의 경우 오차가 라운드에 따라 줄고 있느나, 쉬운 문항에서는 라운드가 진행될수록 오차가 점점 커지는 양상을 보였다. 보통학력 설정에서의 평정자 간 의견 수렴 정도는 문항이 쉬울수록 의견 수렴 정도가 크게 나타났다. 기초학력 설정에서는 라운드에 따른 오차의 변화가 크게 나타나지 않았으며, 모든 라운드에서 문항이 어려울수록 오차가 작았다. 또한, SEE가 라운드에 따라 줄어드는 경향을 보이나 어려운 문항일수록 평정자 간 의견 수렴 정도가 높게 나타났다.



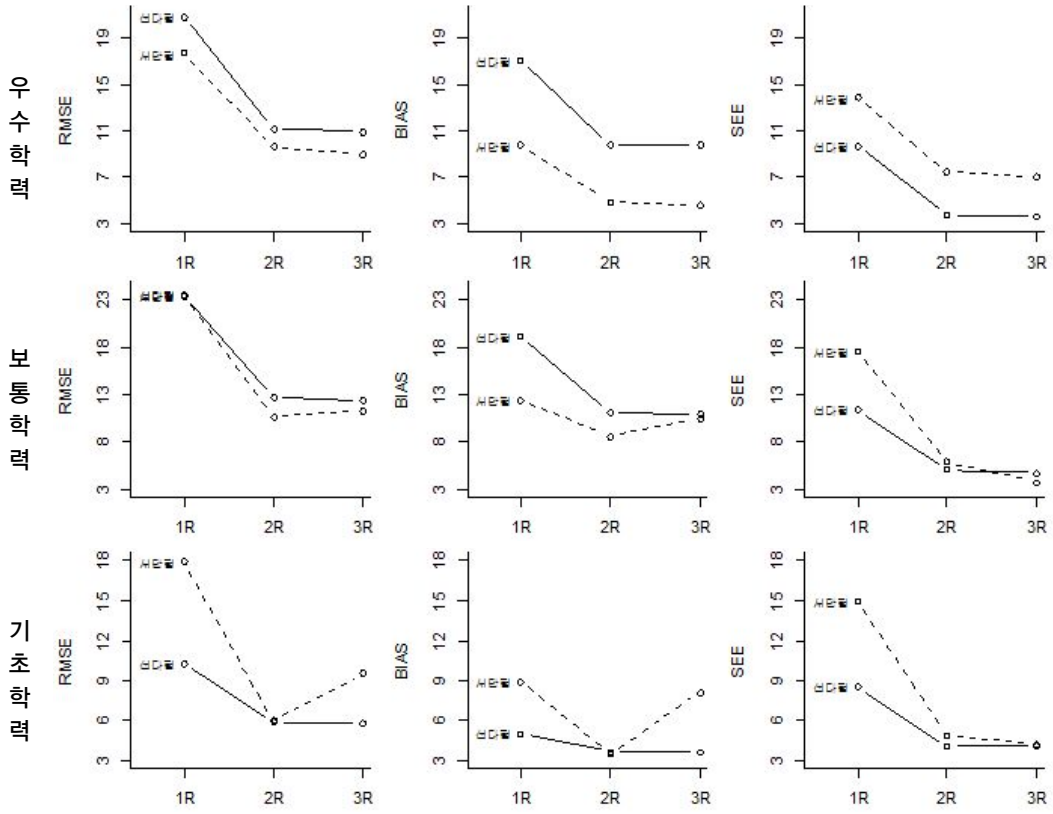
〔그림 3〕 영어 성취수준별 문항 난이도에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

영어의 경우도 수학과 유사한 경향을 보이고 있는데, 우수학력 설정에서는 어려운 문항에서 오차가 상대적으로 컸고, 보통학력 설정에서는 쉬운 문항에서 라운드가 진행될수록 오차가 커진 반면, 난이도 상과 중에서는 줄어들었다. 기초학력 설정에서는 난이도 중인 문항에서 상대적으로 오차가 컸으며, 쉬운 문항에서 가장 오차가 작았다. SEE는 우수와 보통학력 설정에서는 어려운 문항에서, 기초학력 설정에서는 쉬운 문항에서 높았으나 라운드가 진행될수록 줄어들어 최종 라운드에서는 문항의 난이도에 따른 차이가 크지 않았다.

3. 문항 유형에 따른 기대 정답률 예측 오차

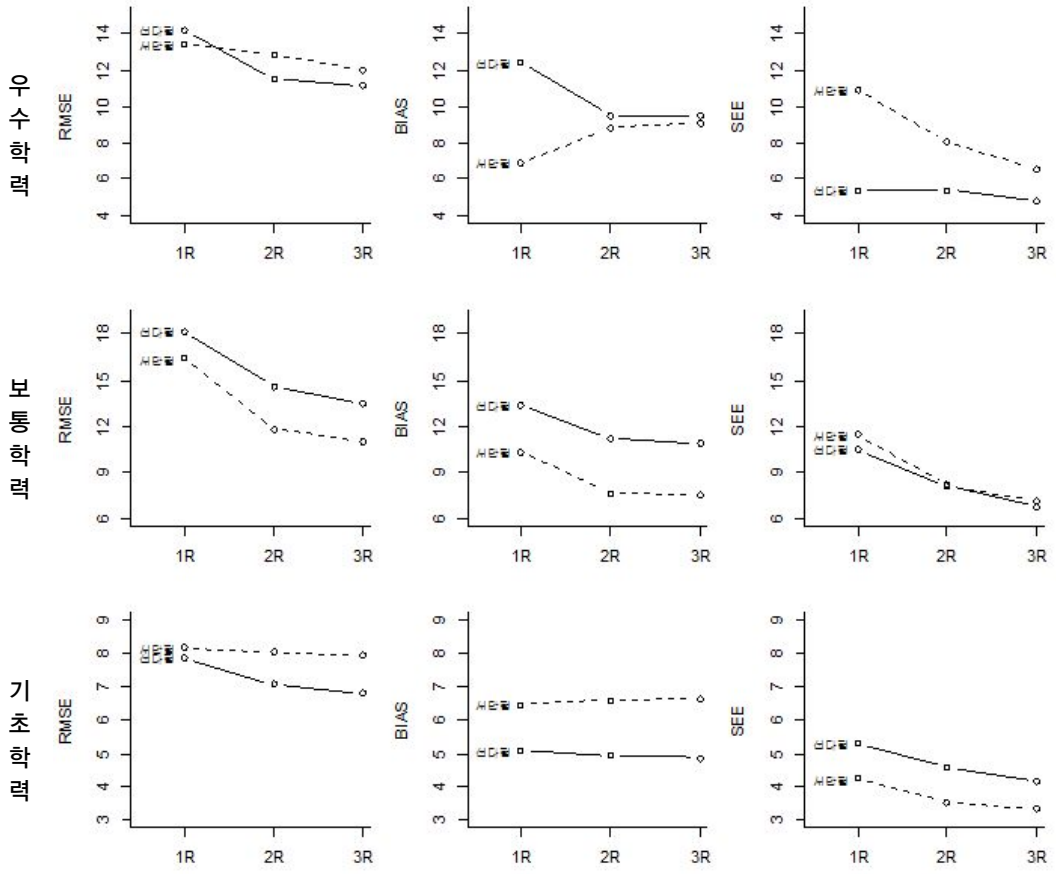
수준설정 과정에서 기대 정답률 오차의 라운드별 변화가 문항 유형 즉, 선다형 문항과 서답형

문항에 따라 차이가 나타나는지를 탐색하였다. 교과별 문항 유형에 따른 기대 정답률의 RMSE, BIAS, SEE는 〔그림 4〕~〔그림 6〕와 같다.



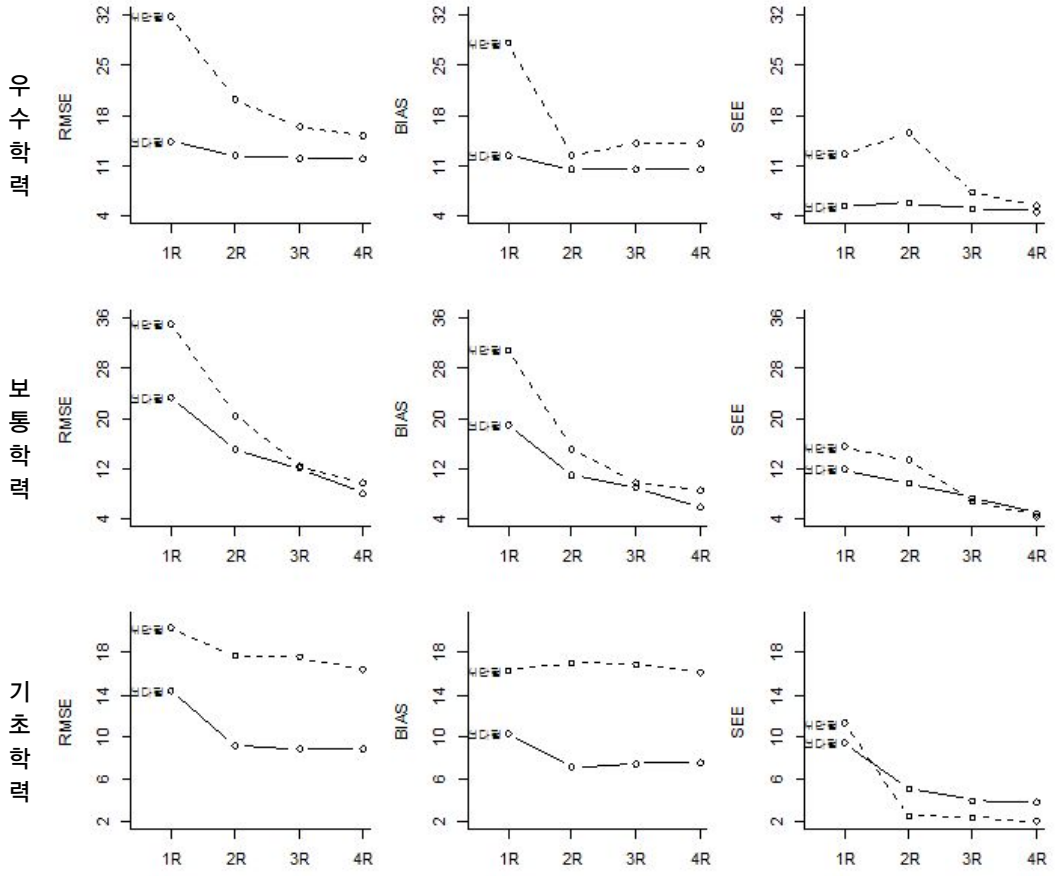
〔그림 4〕 국어 성취수준별 문항 유형에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

문항 유형별 라운드에 따른 기대 정답률의 오차와 평정자 간 기대 정답률의 의견 수렴 정도를 교과별로 살펴보면, 국어의 경우 우수학력 설정에서 선다형 문항의 기대 정답률 예측 오차가 크게 나타났으나, SEE는 서답형 문항이 크게 나타나 평정자 간 의견 수렴 정도는 선다형 문항이 보다 높은 것을 알 수 있다. 보통학력 설정의 경우 우수학력 설정과 유사한 경향을 보이고 있는데, 우수학력 설정에서 라운드에 따른 오차 감소 경향이 보다 크게 나타나 최종 라운드에서는 문항 유형 간 오차 정도와 평정자 간 수렴 정도가 유사하였다. 반면 기초학력 설정에서는 선다형 문항의 경우 서답형 문항보다 오차가 작게 나타나며, 라운드에 따라 감소 양상을 보여 오차가 미미하였다. 기초학력 설정에서의 SEE는 서답형 문항에서 크게 나타났으나 선다형 문항보다 라운드에 따른 감소 정도가 크게 나타나 최종 라운드에서는 문항 유형 간 큰 차이를 보이지 않았다.



〔그림 5〕 수학 성취수준별 문항 유형에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

수학의 우수학력 설정에서는 선다형 문항의 기대 정답률 예측 오차가 라운드에 따라 감소하는 양상을 보이고, 평정자 간 차이도 매우 작게 나타나는 반면, 서답형 문항에서는 오차가 라운드에 따라 증가하는 양상을 보이고 SEE도 라운드에 따라 감소하나 선다형 문항에서 보다 크게 나타난 것을 볼 수 있다. 보통학력 설정에서는 전반적으로 선다형 문항에서 오차가 상대적으로 크게 나타났으나, 평정자 간 기대 정답률 의견 수렴 정도는 문항 유형 간 큰 차이를 보이지 않았다. 기초학력 설정에서는 서답형 문항이 라운드에 따라 큰 변화 없이 상대적으로 오차가 크게 나타났으나, SEE는 선다형 문항이 서답형 문항보다 크게 나타나 서답형 문항에서 평정자 간 기대 정답률의 의견 수렴 정도가 높게 나타났다.



〔그림 6〕 영어 성취수준별 문항 유형에 따른 기대 정답률 예측의 RMSE, BIAS, SEE

영어에서는 우수와 보통학력 설정에서 라운드에 따른 오차 양상이 유사하게 나타났는데, 라운드 초반에는 서답형 문항의 오차가 상대적으로 크게 나타났으나, 라운드가 진행됨에 따라 오차의 감소 정도가 크게 나타나 최종 라운드에서는 문항 유형 간 오차 정도가 유사하였다. 기초학력 설정에서는 서답형 문항에서 라운드에 따른 큰 변화 없이 오차가 상대적으로 크게 나타났다. SEE의 경우 전반적으로 서답형 문항이 크게 나타났으나, 라운드가 지남에 따라 오차가 줄어들어 최종 라운드에서는 문항 유형 간 큰 차이를 보이지 않았다.

Ⅴ. 결론 및 논의

앙고프 방법을 적용한 수준설정 과정에서는 최소능력자에 대한 개념화나 기대 정답률의 예측 등에 있어 전문가들의 주관적인 판단이 개입될 가능성이 있다. 이러한 가능성으로 인해 평정자들이 수준설정 과정에서 최소능력자 집단으로 개념화한 것이 기대 정답률을 통해 성취수준의 실제로 발현될 때, 그 특성이 정확히 반영되지 않을 수 있다. 이 연구는 수준설정에 대해 절차적 측면의 타당성 확인을 목적으로, 변형된 앙고프 방법에서 평정자의 기대 정답률 예측에 대한 정확성과 문항의 특성에 따른 기대 정답률 예측의 오차를 분석하였다. 수준설정에서의 최소능력자의 기대 정답률과 수준설정을 통해 산출된 분할점수 집단의 실제 정답률의 차이를 통해 기대 정답률 판정의 예측 정확성 지수를 산출하였으며, 문항의 난이도와 문항 유형에 따라 라운드별 기대 정답률의 오차 경향을 탐색하였다. 주요 연구 결과를 중심으로 결론과 논의를 제시하면 다음과 같다.

첫째, 라운드에 따른 기대 정답률의 오차는 교과 및 성취수준별로 차이가 있었는데, 모든 교과에서 보통학력에 대한 수준설정 시 기대정답률의 오차가 가장 크게 나타났다. 또한, 수학 교과의 경우는 라운드에 따른 오차 정도가 변화가 없었음을 확인할 수 있었다. 이러한 점은 성취수준에서의 양 극단에 위치하고 있는 우수학력과 기초학력의 특성이 보통학력의 특성보다 더 명확해서 나타나는 현상으로 이해된다. 즉, 보통학력인 학생 중 최소의 능력을 지닌 학생의 개념화를 보다 뚜렷이 해야 할 필요성을 보이며, 보통학력의 학생을 구분하기 위한 수준설정 과정에서 기초학력과 보통학력의 특성 차이를 보다 분명하게 해야 할 필요성을 보여주며, 이를 염두에 두고 평정자 훈련이 진행되어야 함을 시사한다. 특히, 성취평가제가 본격적으로 시행되고 있는 시점에 성취수준을 보다 정교하게 설정하기 위해서는 각 교과별 가장 높거나 가장 낮은 수준에 비해 가운데 수준의 학생 특성을 더 명확하게 정의하고, 평정자가 해당 성취수준을 명확히 개념화 할 필요가 있다.

둘째, 국어, 수학, 영어 모든 교과의 모든 수준설정에서 라운드가 진행됨에 따라 평정자 간 차이가 크게 줄어들어 평정자들의 라운드별 최소능력자 및 기대 정답률 예측 과정의 타당성을 확인할 수 있었다. 특히 학생들의 평균 수행수준에 대한 경험적 자료를 제공하여 이를 기반으로 평정자 간 논의가 이루어졌던 2라운드에서 오차가 크게 감소하였으며, 이후 라운드에서도 지속적으로 오차가 감소하여 평정 라운드가 진행됨에 따라 평정자 간 최소능력자의 논의에 따른 의견이 점차 수렴되었음을 확인하였다. 또한, 평정자 간 문항 정답률에 대한 논의가 이루어지지 않은 상태인 1라운드에서의 기대 정답률은 실제 해당 성취수준에 포함되는 모든 학생의 정답률을 기준으로 할 경우와 유사한 것을 확인할 수 있었는데, 이는 1라운드에서 평정자들이 최소능력자를 대부분 성취수준별 최소능력 수준의 학생이 아닌 평균적인 능력의 학생으로 개념화한 것

을 알 수 있었다.

셋째, 문항의 난이도에 따른 라운드별 기대 정답률의 오차는 지속적으로 감소하고 있으나, 문항이 매우 어렵거나 쉬운 문항에서 기대 정답률의 오차가 상대적으로 크게 나타났다. 문항 난이도에 따른 오차 결과는 Impara와 Plake(1998)의 연구에서도 동일하게 나타났는데, 이 연구에서는 모든 교과의 우수학력과 기초학력의 수준설정에서 두드러졌다. 특히 난이도 상, 즉 어려운 문항의 경우 평정자들이 기대 정답률을 과대 추정하는 경향을 보이며, 쉬운 문항의 경우 과소 추정하는 경향을 보이고 있었다. 즉, 실제 정답률이 높은 문항에 대해서는 기대 정답률을 보다 낮게 추정하고, 실제 정답률이 낮은 문항에 대해서는 기대 정답률을 보다 높게 추정하였다. 이는 평정자들의 라운드별 최소능력자에 대한 기대 정답률 예측에 문항의 난이도가 영향을 미치며, 이러한 오차로 인해 산출된 분할점수가 실제 평정자들이 개념화한 최소능력자의 수행수준을 반영하지 못할 가능성을 보여준다.

특정한 목적의 진단검사와 같이 대부분의 문항이 쉽거나 어려운 문항으로 구성되어 있을 경우, 수준설정에 보다 큰 영향을 미칠 수 있다. 대부분 쉬운 문항으로 구성된 검사의 경우, 평정자의 기대 정답률에 대한 과소 추정으로 인해 해당 성취수준에서 평정자가 기대하는 최소능력자의 획득점수를 과소 추정하게 되어 분할점수가 낮게 설정될 수 있으며, 어려운 문항으로 구성된 검사의 경우에는 평정자의 기대 정답률에 대한 과대 추정이 분할점수의 과대 추정으로 이어져 평가 결과의 잘못된 해석을 유도할 수 있다. 또한, 평정자들의 의견 수렴 정도는 문항의 난이도에 따라 큰 차이를 보이고 있지 않아 대부분의 평정자들이 전반적으로 기대 정답률 예측에 오차가 개입될 수 있다. 이러한 점은 수준설정 과정에서 문항 난이도에 따른 평정자들의 기대 정답률 예측 훈련의 필요성을 보여주며, 라운드별 기대 정답률 예측의 평정자 간 논의에 문항의 난이도를 고려하여 기대 정답률 예측의 정확성을 높여야 할 필요성을 보여준다.

넷째, 문항 유형에 따른 라운드별 기대 정답률의 예측 오차는 존재하며, 전반적으로 기초학력 수준설정 시 서답형 문항에서의 오차가 상대적으로 크게 나타났으나, 교과와 성취수준별로 서로 다른 경향을 보이고 있었다. 이러한 결과는 문항 유형이 수준설정 과정에서 평정자가 기대 정답률을 추정하는 데에 영향을 미치나 그 효과는 교과, 성취수준에 따라 다르게 나타나는 것을 의미한다. 문항 유형에 따른 정답률 예측 과정을 보면, 서답형 문항은 선다형 문항과는 다르게 하위 문항별로 각 부분 점수에 대해 기대 정답률을 예측하며 부분 점수에 대한 기대 정답률 예측의 합은 100%가 되도록 하였다. 이러한 절차는 서답형 문항에 대한 수준설정 시 일반적으로 적용되는 과정인데, 이 과정 속에서 평정자들이 성취수준별 기대 정답률을 예측할 때, 마지막 부분 점수에 대한 예측은 이전에 예측된 결과에 따라 자동적으로 결정된다. 또한, 각 교과, 성취수준별 서답형 문항에 대한 평정자들의 기대 정답률 예측에 있어 중점적으로 예측하는 부분 점수가 다를 수 있어, 교과, 성취수준별 문항 유형에 따른 예측 정확성 정도의 차이를 발생시킬 수 있다. 이러한 결과는 문항 유형에 따라 교과, 성취수준에 따른 평정자의 정답률 예측 훈련 및 라운드별 논의를 다르게 진행해야 할 것을 제안하며, 특히, 서답형 문항의 기대 정답률 예측

에 있어 부분 점수에 대한 평정자들 간의 논의가 필요하다고 지적할 수 있다.

이 연구에서는 문항의 특성인 문항 난이도와 문항 유형에 따른 성취수준 설정에서의 예측오차를 점검하고 있는데, 성취수준 설정 과정에서의 오차에는 문항에 대한 특성뿐 아니라 평정자의 특성이 영향을 미칠 수 있다. 수준설정 과정에서의 오차를 최소화하기 위해서는 평정자의 특성이 기대 정답률 예측에 미치는 영향을 탐색해야 할 필요가 있으며, 이러한 결과를 바탕으로 성취수준을 보다 타당하고 정교하게 설정해야 할 것이다.

수준설정 과정에서 평정자들의 기대 정답률 예측 훈련과 라운드별 논의를 통해 정답률 예측의 정확성을 높이는 것은 성취수준 과정에서 평정자들이 기대하는 최소능력자를 분할점수에 보다 정확하게 반영할 수 있도록 유도한다. 학업성취도 평가뿐만 아니라 대학수학능력시험의 일부 영역에 준거참조평가가 도입되고, 일선 학교에서 성취평가제가 시행됨에 따라 교사에 의한 평가 결과도 준거참조평가 형태로 보고된다. 최근 들어 교육적 가치를 위해 더 강조되는 형성평가의 그 의미와 취지를 살리려면 준거참조평가가 적용되어야 한다. 대규모 평가이든 교실 평가이든 준거참조평가에 의한 결과를 얻기 위해 수준설정의 중요성과 필요성이 강조되고 있다. 이 연구의 결과를 기반으로 하여 수준설정의 평정자 훈련 및 정답률 예측 과정의 고려 사항을 통해 실제 성취수준의 최소능력자 특성이 수준 설정에 정확하게 반영되도록 해야 할 것이다.

참 고 문 헌

강태중(2014). 대학수학능력시험 발전 방향 모색. 제26회 KICE 교육과정·평가 정책포럼 수능 영어영역 절대평가 도입 방안 탐색. 한국교육과정평가원 연구자료 CAT 2014-13.

강태훈, 박찬호, 김인숙(2011). 검사 길이와 수행등급 개수에 따른 성취수준 분류일관도 및 정확도 연구. **교육평가연구, 24**(4), 1017-1038.

교육부(2013). 2017학년도 대입제도 확정 발표 보도자료(2013. 10.24).

김석우, 윤명희, 지은림(1999). 준거지향검사 기준설정 방안의 비교분석. **교육학연구, 37**(2), 227-247.

김성숙, 송미영, 박인용(2012). 다변량 일반화가능도 이론을 적용한 성취수준 설정에서의 오차분석과 최적 조건 탐색. **교육평가연구, 25**(4), 581-602.

김성숙, 송미영, 최인봉, 김희경(2010). 2010년 국가수준 학업성취도 평가 기술보고서. 한국교육과정평가원 연구보고 RRE 2010-7-5.

김신영(2009). 대학수학능력시험의 개선 방안 탐색. **교육평가연구, 22**(1), 1-27.

민경석(2013). 한국과 미국 국가수준 학업성취도 평가의 성취수준 비교. **중등교육연구, 61**(1), 111-136

박연복, 이규민, 강상진(2011). 군집분석을 이용한 수준설정 방법과 타당성 연구. **교육평가연구, 24**(3), 645-664.

박정, 김경희, 김수진, 손원숙, 송미영, 조지민(2006). 국가수준 학업성취도 평가: 기술보고서. 한국교육과정평가원 연구보고 RRO 2006-4.

백순근, 최인희(2006). 준거지향평가 기준설정을 위한 Rasch 방법의 숙달 학습자판정 일치도:원점수 및 Angoff 방법과의 비교를 중심으로. **교육평가연구, 19**(2), 157-178.

송미영, 김성숙, 박인용(2013). 대규모 준거참조평가에서 성취수준의 분류 일치도와 정확도 분석. **교육평가연구, 26**(2), 391-413.

양길석(2000). 준거지향평가의 준거설정 방법 비교: 중학교 논술 기초능력 검사를 중심으로. **교육평가연구, 13**(2), 107-133.

이규민(2004). 의사 국가시험 합격선 설정에 관한 측정학적 접근. **보건의료교육평가, 1**(1), 5-14.

이인제, 최석진, 이재기, 이봉주, 채선희, 김도남, 강미현, 김혜숙, 이규민, 김수정(2004). 2003년 초등학교 3학년 국가 수준 기초 학력 진단 평가 연구: 종합. 한국교육과정평가원 연구보고 CRE 2004-1-1.

이현숙(2007). 반복측정의 관점에서 본 분류일치도 계수의 유형. **교육과정평가연구, 10**(1). 103-119.

장윤선, 성태제(2009). 문항반응이론에 기초한 준거설정 방법 비교. **교육평가연구, 22**(3), 659-680.

한국교육개발원(2014). 제 63차 KEDI 교육정책 포럼 - 수능 영어 과목 절대평가 도입 공개 토론회 자료집. 한국교육개발원 연구자료 RRM 2014-01-3.

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

한국교육과정평가원(2014). 성취평가제 적용, 이렇게 하세요 - 고등학교 보통교과용 -. 한국교육과정평가원 연구자료 ORM 2014-20.

한국교육평가학회(2004). **교육평가 용어사전**. 서울: 학지사.

Allen, N. L., Carlson, J. E., Johnson, E. G., & Mislevy, R. J. (2001). Scaling procedures. In N. L. Allen, J. R. Donghue, & T. L. Schoeps (Eds.), *The NAEP 1998 technical report, NCES 2001-509,* 227-246. U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Angoff, W. B. (1971). Norms, Scales, and equivalent scores. In R. L. Thorndike (Ed.). *Educational Measurement*(2nd Ed.). Washington, D. C.: American Council on Education.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*(2), 219-249.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15*(2), 20-31.

Cizek, G. J., & Bunch, M. (2007). *Standard setting.* California: Sage Publication, Inc.

Giraud, G. T., Impara, J. C., & Buckendahl, C. W. (2000). Making the cut in school districts: Alternative methods for setting cut-scores. *Educational Assessment, 6,* 291-304.

Glass, G. V. (1978). Standard and criteria. *Journal of Educational Measurement, 15,* 237-261.

Impara, J. C., & Plake, B. S. (1998). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366.

Jeager, R. M. (1992). Establishing standards for teacher certification test. *Educational Measurement: Issues and Practice, 9,* 15-20.

Kane, M. T. (1986). *The interpretability of passing scores*(Tech. Bulletin No. 52). Iowa City, IA: American College Testing Program.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

Kane, M. T. (2001). So much remains the same: Conception and Status of Validation in Setting Standards. In: Cizek G. J., ed. *Setting performance standards: concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 53-88.

Lee, G., Park, I.-Y., Lee, M.-S., Park, Y., & Kim, K.-S. (2010). *The validity of the Angoff and Bookmark Standard Setting Methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24*, 56-64.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek(Ed.), *Standard setting: Concepts, methods, and perspectives,* 119-157. Mahwah, NJ: Erlbaum.

Shepard, L. A. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Ed.), *Practices and problems in competency-based measurement*. Washington, DC: National Council on Measurement in Education.

Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.

van der Linden, W. J. (1995). *A conceptual analysis of standard setting in large-scale assessments*. In Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), II, 97-117. Washington, DC: U. S. Government Printing Office.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek(Ed.), *Standard setting: Concepts, methods, and perspectives,* 19-51. Mahwah, NJ: Erlbaum.

• *논문접수 : 2014-08-29/ 수정본접수 : 2014-09-30/ 게재승인 : 2014-10-13*

ABSTRACT

Accuracy of estimating expected probability for MAP in standard setting

In-Yong Park
(Associate Research Fellow, Korea Institute for Curriculum and Evaluation)
Mi-Young Song
(Research Fellow, Korea Institute for Curriculum and Evaluation)
Sungsook Kim
(Senior Research Fellow, Korea Institute for Curriculum and Evaluation)

This study investigated procedural validity of NAEA(National Assessment of Educational Achievement) standard setting by examining an accuracy of expected probability for MAP(minimally acceptable person). The expected probability for MAP obtained from standard setting procedure in NAEA were compared with real proportion for students who earned cut-score at each achievement level in terms of item difficulty and type. As a results, we confirmed the procedural validity of standard setting in NAEA based on the fact that the magnitude of overall difference between expected probability and real proportion were decreased as a round was progressed. We also found panels overestimated the correct probability in difficult items and underestimated the correct probability in easy items. The magnitude of accuracy in item type showed different patterns across subjects and achievement levels. We also suggested to train the panel in order to obtain more authentic cut scores in standard setting based on the results of this study.

Key Words : Standard setting, Angoff method, Expected probability, Procedural validity