

수학 지필 평가의 실제 분석

서 동 업¹⁾

본 연구는 수학 지필 평가를 효과적으로, 그리고 평가의 목적을 구현하기 위해 적절한 방법으로 시행하기 위하여 해결해야 할 문제를 알아보고 이에 대한 해결 방안을 탐색해 보고자 한 것이다. 이를 위하여 주로 교육성취도 평가 연구와 제 3차 수학·과학 성취도 국제 비교 반복 연구, 영재 판별 검사 등 최근에 이루어진 평가의 실재를 분석하였다. 분석의 기준은 평가 틀, 결과 해석, 문항의 제시 형식에 따른 문항 곤란도, 국제 평가 문항의 번역 등의 네 가지이다. 평가 틀과 관련하여 최근 들어 강조되고 있는 통합적인 성격의 문항에 대한 분류가 필요하다는 점을 주장하였으며, 결과 해석과 관련하여 전형적인 이원 분류에서 이용되는 행동 영역에 대한 결과 해석이 그리 잘 이루어지지 않고 있어 결과 해석에서 의미를 줄 수 있는 평가 틀에 대한 고려가 필요함을 주장하였다. 또한, 문항 곤란도를 결정하는 여러 조건과 소재에 대한 문제를 논함으로써 교육 과정의 성취도를 평가할 때 보다 명시적인 기준에 대한 논의가 필요함을 주장하였으며, 국제 평가에서 번역 과정에서 생길 수 있는 문제를 지적함으로써 이에 대한 신중한 연구가 필요함을 주장하였다.

I. 서론

수학교육에서 지필 평가는 다양한 장면에서 이용되고 있다.²⁾ NCTM에서는 종래에 지필 검사만 이용하던 것에서 탈피하여 지필 검사, 구두시험 및 시범 등 다양한 평가 기법을 사용할 것을 권고하고 있으며(구광조 외 공역, 1992), 우리 나라의 수학교육에서도 수행평가로 대표되는 다양한 평가 방법의 필요성이 대두되고 있기는 하지만, 대부분의 평가 상황에서 지필 평가는 여전히 중요한 역할을 하게 될 것으로 보인다. 이러한 지필 평가로는 먼저, 1999년에 시행되어 2000년에 결과가 발표된 제 3차 수학·과학 성취도 국제 비교 반복연구(이하 TIMSS-R로 표기)나 최근에 결과가 발표된 OECD 국가간 학업성취도 국제 비교 연구, 해마다 시행되는 국제 고등학생 수학 올림피아드 등과 같은 국제적인 규모의 국가간 비교 평가를 들 수 있다. 또한, 전국 규모로 이루어지는 대학수학능력시험이나 교육성취도 평가, 각종 경시대회, 학교 단위의 중간시험이나 기말시험, 교사의 수업 단위의 진단평가나 형성평가, 교사의 관찰 평가, 학생들의 자기 평가 등에 이르는 국내 평가를 들 수 있다. 이렇듯 다양한 유형의 지필 평가가 다양한 장면에서 이용되면서 다양한 기능을 수행하고 있는 것이다.

이러한 각각의 평가는 고유의 목적을 지니고 있다. 이를테면, Robitaille et al.(1993)은 TIMSS를 통하여

1) 춘천교육대학교([200-703] 강원도 춘천시 석사동 339)

2) Niss(1993)는 평가에 대한 영어 용어로서 'assessment'는 다양한 학생들의 개인적 또는 집단적 수학적 능력이나 성취도를 판단하기 위해 실시하는 것으로, 'evaluation'은 전체적 또는 부분적으로 교육제도나 수업 체계를 판단하기 위해 실시하는 것으로 구분하고 있다. 본 고에서 다루는 평가는 Niss의 용어 구분에 따른다면 'assessment'에 해당하는 것이다.

답하고자 하는 문제로 크게 다음의 두 가지를 들고 있다. 첫째는 국가간에 수학 및 과학 교육의 환경은 얼마나 다른가, 그리고 학생들의 결과는 얼마나 다른가, 결과의 차이는 수학 및 과학 교육 환경의 차이와 어떻게 관련되는가의 문제이다. 둘째는 수학 및 과학 교육의 이론과 실제의 개선으로 이끌 수 있는 문맥과 투입, 결과 사이의 관계에 대한 패턴이 존재하는가의 문제이다. 이러한 목적에 따른다면 우리는 TIMSS-R의 수학 분야에서 2등이라는 우수한 성적을 거두었다는 것도 중요하겠지만, 이와 더불어 우리나라 학생들이 우수한 성취를 보인 내용 영역이나 문항의 경우 어떤 원인으로 인하여 그런 결과를 거두었는지를 분석하는 일이 중요하다.³⁾ 또한 우리 나라보다 다른 나라의 학생들이 우수한 성취를 보인 영역이나 문항이 있다면 그 나라의 교육 과정이나 지도 방법은 어떠한지를 살펴보는 일이 중요할 것이다.

그리고, 대학수학능력시험이라면 고등학교에서 학습한 내용을 소재로 하여 장차 대학에 진학하여 학업을 진행할 수 있는 능력을 측정하는 평가이므로, 이 목표에 적합하기 위해서는 시험에서 성취도가 높은 학생일수록 대학에 진학하여 학업을 잘 수행할 수 있는 문항이 출제되어야 할 것이다. 아울러 수학 능력 시험에서 우수한 성취를 보인 학생이 대학에서도 우수한 학업 성취를 보이고 있는지에 대한 치밀한 사후 연구도 필요한 것으로 보인다. 또한, 전국의 표집 학교에 대하여 시행되는 교육성취도 평가 연구의 주요 목적의 하나는 교육 과정의 질 관리에 있으므로(김명숙 외, 1999), 교육 과정의 내용을 충실히 반영하는 문항을 출제하는 것이 기본적으로 전제되어야 할 중요한 문제이다.

제 7차 교육과정의 시행과 더불어 수학교육에서 정교한 평가의 중요성은 더욱 커지고 있는 것으로 보인다. 제 7차 교육과정에서 수학 교과는 단계형 수준별 교육과정으로 운영되며, 학교 단위의 교육 과정의 운영 지침에서 '단계형 수준별 교육 과정 운영에 있어서, 교과, 학년 또는 단위 학교의 상황에 따라 해당 단계 학습 목표의 일정한 성취 기준을 고려하여 차상급 단계의 진급을 위한 자격 기준을 설정한다'고 명시하고 있다(교육부, 1997a : 21). 그렇기에 어느 단계의 수업을 받은 학생들이 차상급 단계로의 진급을 위한 자격 기준을 충족시키고 있는지의 여부를 판단하기 위하여 그 구체적인 형태야 어떻든 간에 모종의 평가가 이루어질 수밖에 없기 때문이다. 또한, 이 경우에 어떤 문항에 대하여 어느 정도의 성취를 보인 학생을 차상급 단계로 진급시킬 수 있는지에 대한 기준을 마련하는 문제도 상당한 논의와 연구가 필요할 것으로 보인다.

본 연구의 목적은 이렇듯 다양한 장면에서 이루어지고 있는 지필 평가의 실재를 분석함으로써, 주어진 평가 목적을 구현하기 위하여 필요한 지필 평가 방법의 기초에 대한 여러 문제를 탐색하려는 것이다. 이를 위하여 한국교육과정평가원이 주관이 되어 시행하고 있는 교육성취도 평가와 최근에 실시되어 우리나라가 종합적으로 2위를 차지했던 TIMSS-R, 한 종류의 영재 판별 검사를 주요 소재로 하여, 크게 두 가지의 문제를 다루고자 한다. 첫째로, 평가를 실시하는 과정에서 이용하게 되는 평가 틀과 이에 따른 평가 결과 해석의 실재를 분석하고자 한다. 수학 교과를 비롯하여 대부분의 평가에서는 문항을 제작하는 과정에서 Tyler에서 유래하는 이원분류를 활용하고 있다. 이러한 이원분류가 활용되는 실재를 조사하고 주어진 평가의 목적과 평가 틀, 이에 따른 결과 해석 과정을 분석하여, 원래 평가의 목적을 구현하는 데 적절한 방향을 모색하고 이를 위해 필요한 일을 찾아보고자 한다. 둘째로, 여러 평가에서 이용된 문항과 그 결과의 실재를 분석한다. 이를 위하여 여러 장면에서 이용되고 있는 평가 문항과 그에 대한 결과를 조사하며, 그럼으로써 주어진 평가 목적에 적합한 문항을 제시하는 데 있어 유발되는 여러 어려움과 이를 해소

3) 본 연구에서는 지필 평가에 이용되는 도구를 지칭하기 위하여 '문제(problem)'라고 하는 대신 '문항(item)'이라는 용어를 이용한다. 강문봉 외 공역(1999)에서는 다음과 같이 '문제'의 의미를 밝히고 있다. '문제는 누군가 무엇인가를 원하고 있지만 즉각 그것을 얻기 위해 무엇을 해야 할지 모르는 상황과 관련 있다.' 수학교육의 평가 장면에서 이용되는 도구는 대다수 학생이 쉽게 알 수도 있는 기본적인 내용이라고 하더라도 교육과정에 제시된 내용이라면 평가해야 할 상황이 있다. 그렇기에 본 고에서는 용어의 혼란을 피하기 위하여 '문항'이라는 용어를 이용하려는 것이다.

하기 위한 아이디어를 얻는 것을 목적으로 한다.

II. 지필 평가의 틀과 결과 해석의 실제

제 7차 수학과 교육과정에서 수학과목의 목표는 ‘수학의 기본적인 지식과 기능을 습득하고, 수학적으로 사고하는 능력을 길러, 실생활의 여러 가지 문제를 합리적으로 해결할 수 있는 능력과 태도를 기른다’고 규정되어 있다(교육부, 1997a : 29). 이러한 목표에서 드러나듯이 이후에 평가가 가능한 행동적인 형태로 교육 목표를 서술하는 것이 Tyler 이래로 하나의 중요한 교육적 전통이 되어 온 경향이 있다. 본 절에서는 수학과목의 평가 장면에서 Tyler의 전통에서 유래되어 활용되고 있는 평가 틀과 이를 이용한 평가 결과의 해석에 대한 실례를 분석해 보아 평가자의 목적을 구현하는 평가를 시행하는 과정에 도움을 줄 수 있는 방안을 모색해 보기로 한다.

1. 수학 교육에서 평가 틀의 실제

Tyler의 제안에 의하면 교육 목표는 ‘어떤 내용에 관한 어떤 행동’이라는 식으로 ‘내용’과 ‘행동’을 이원적으로 표시하는 것이어야 한다(이홍우, 1995). 예컨대, 이차방정식을 지도한다면 Tyler의 입장에서 목표 진술은 ‘이차방정식을 풀 수 있다’는 식으로 표시되게 되지만, 궁극적으로 이차방정식을 푸는 것이 왜 가치가 있는지의 문제는 교육 철학의 문제로 이미 이차방정식을 지도할 가치가 있는 것으로 가정하고 있다.⁴⁾ 그 동안 국내외에서 이루어진 여러 평가 장면에서 Tyler식의 교육 목표의 이원적 진술에 근거한 이원분류는 하나의 주요한 전통이 되어 온 것으로 보인다.

Tyler의 전통을 이은 Bloom은 행동 영역을 크게 인지적 영역과 정의적 영역으로 나누고, 인지적 영역을 1.00 지식, 2.00 이해력, 3.00 적용력, 4.00 분석력, 5.00 종합력, 6.00 평가력의 6개 항목으로, 정의적 영역을 1.00 감수, 2.00 반응, 3.00 가치화, 4.00 조직화, 5.00 인격화의 5개 항목으로 나누고 있다(이홍우, 1995).⁵⁾ 그러나, 김응태 외(1984)에서 지적하고 있듯이 Bloom의 이러한 행동 수준은 수학과목에 그대로 적용하기 모호하다는 점 때문에 이를 수학과목에 맞게 수정하고자 한 여러 시도가 있었다. 이를테면 $27 + 68$ 이라는 문제는 덧셈의 받아올림에 대한 지식 문제로 볼 수도 있으며, 받아올림이 있는 덧셈 방법에 대한 이해 문제로 볼 수도 있으며, 십진기수법의 원리를 적용하는 문제로 볼 수도 있다는 것이다. 최근에 시행되고 있는 여러 평가의 행동 영역을 보더라도 Bloom의 구분을 그대로 적용하는 예를 찾기는 쉽지 않으며, 오히려 평가마다 다양한 행동 영역이 이용되고 있음을 알 수 있다.

먼저 살펴보고자 하는 평가의 예는 교육성취도 평가로서, 여기서는 수학의 평가 틀로서 내용 영역과 행동 영역이라는 이원분류를 이용하고 있다. 내용 영역은 제 7차 수학과 교육과정 상의 6개 영역에 따라 구분하고 있으며, 행동 영역은 계산, 이해, 추론, 문제해결, 의사소통의 5개 영역으로 구분하고 있다(김명숙

4) 바로 이 점 때문에 이홍우(1995)는 Tyler 식의 목표 진술보다는 교육 내용 자체로부터 교육의 목표와 구체적인 지도 내용을 정당화하는 Bruner의 입장을 지지하고 있다. 본 연구자가 교육 목표에 관한 Tyler의 입장에까지 동의하는 것은 아니지만, 무엇보다도 본 연구는 수학교육의 다양한 장면에서 실제로 활용되는 평가에 관한 것이며, 이원분류가 주요한 평가 장면에서 이용되고 있기에 본 연구에서 다루고자 하는 것이다.

5) 본 연구에서 다루고자 하는 평가는 주로 지필 평가에 국한되며, 이러한 평가 장면에서 정의적 행동 영역을 다룬 경우를 찾아보는 것은 힘들다. 따라서, 행동 영역에 대한 문제는 주로 인지적 행동 영역에 대한 것으로 제한된다.

외, 1999). 여기서, 계산, 이해, 추론, 문제해결의 4개 영역은 우정호(1998)가 구분한 수학적 사고의 4개 영역과 일치하는 것이며,⁶⁾ 의사소통 영역은 계산, 개념, 추론, 또는 문제해결 영역에 관한 문제를 해결하는 상황에서 주어진 문제 상황과 관련된 수학적 내용을 토대로 수학적 용어, 기호, 문장 등을 이용하여 그 해결 과정의 근거 및 이유를 표현할 수 있는 능력에 관한 것으로 정의하고 있다.

이처럼 내용 영역은 교육과정의 영역으로, 행동 영역은 수학적 사고력을 기준으로 나누는 것은 학생들의 학업 성취를 판단하는 많은 평가에서 활용되는 방법이다. 하지만, 이러한 내용 영역과 행동 영역의 구분에 대하여 몇 가지 문제를 제기해 볼 수 있다. 먼저 내용 영역과 관련하여 각각의 내용 영역이 포함하는 범위에 대한 문제를 생각해 볼 수 있다. 예를 들어, TIMSS-R에 출제되었던 문항 중 '400L의 물을 채우는 데 250mL의 병이 모두 몇 개 필요한가?'(김성숙 외, 1999)라는 문항은 들이의 단위와 관련된다는 점에서 측정 영역에 포함되며, $400 \div 0.25$ 든 아니면 $400000 \div 250$ 이든 다소 복잡한 나눗셈을 요구한다는 점에서 수와 연산 영역에도 포함된다.

이와 같은 통합적인 문항은 NCTM(1989)에서 평가할 항목의 하나로 '수학적 힘'을 제시하면서 통합적 능력을 제시한 것과 맥을 같이 하여 여러 평가에서 그 활용도가 집중하고 있는 추세이다. 그리고, 2개 영역 이상에 걸쳐 있는 문항의 경우 내용 영역은 중심적인 아이디어가 포함되어 있는 영역으로 구분하는 것이 일반적으로 보이며, 박정 외(2000)의 연구에서는 TIMSS-R에 출제된 문항을 국내 교육과정에 따라 재분류하면서 통합적인 문항을 묶어 '독립영역'이라고 별도로 분리하기도 하였다. 산술적으로 계산해 볼 때, 2개 영역을 통합한 문항에 대하여 통합된 영역을 한 쌍으로 하여 정확한 영역을 표시한다면 15개의 내용 영역이 추가로 필요하게 된다. 결국 어떤 형식의 내용 영역을 따르며 어느 정도로 세분할 것인가의 문제는 내용 영역의 구분이 평가의 결과의 해석에 어떻게 활용될 것인가의 문제와 같이 생각해야 할 것으로 보인다.

또한, 행동 영역과 관련하여 다음과 같은 우정호(1998)의 지적을 참고할 필요가 있다.

계산법도 개념적 사고도 결국은 문제를 풀기 위한 수단, 도구에 불과하다고 할 수 있다. 수학 문제에는 답을 구하는 문제와 증명 문제가 있다. 수학적 능력의 핵심은 수학적 문제해결 능력이라고 볼 수 있을 것이며, 수학교사의 주요 임무는 학생들의 문제해결 능력을 개발하는 것이다.

이러한 지적에 비추어 본다면 계산 문제나 개념의 이해를 묻는 문제, 증명 문제는 모두 문제 해결에 속하는 것으로 볼 수도 있다. 실제 평가가 이루어지는 장면에서는 계산이나 개념의 이해, 추론을 제외한 수학 내적 문제해결이나 외적 문제해결 문항을 문제해결 영역으로 포함하고 있는 것으로 보인다. 그렇다고 하더라도, '문제해결' 영역이라는 용어로 인한 의미의 오해를 불러일으킬 소지는 있는 것이다. 특히, 평가의 결과를 해석하면서, '계산 능력은 있으나 문제해결 능력은 부족하다'는 식의 결과를 진술하는 것도 의미가 모호해질 위험이 있는 것이다. 또한, '의사소통' 영역의 경우 수학적 용어, 기호, 문장을 이용하여 근거나 이유를 표현하는 능력으로 정의되어 있으나, 이로 인하여 '의사소통' 영역을 측정하는 문항은 서술형 문항으로 제한되게 된다. 이는 서술형 문항을 활용하기 힘든 대규모 평가 상황에서 의사소통 능력을 측정하고자 할 경우 모종의 갈등 상황을 유발할 가능성이 있어 보인다.

6) 이상의 4가지 행동 영역은 대학수학능력시험의 행동 영역이기도 하다(김명숙 외, 1999). 우정호(1998)는 추론을 세분하여 추측, 유추 등의 귀납적 추론과 증명을 의미하는 연역적 추론으로 나누고 있으며, 문제해결을 세분하여 수학적 소재로 한 문제 해결을 뜻하는 내적 문제해결과 다른 학문이나 실생활 상황에 응용되는 측면을 강조하는 외적 문제 해결로 나누고 있다.

국제 평가의 예로서, TIMSS-R에서 이용하는 평가 틀은 내용 영역과 수행(performance) 영역의 이원 분류이다. 내용 영역은 분수와 수 감각, 측정, 자료의 표현 및 해석·확률, 기하, 대수의 5개 영역이며, 행동 영역과 대등한 것으로 보이는 수행 영역에는 지식, 복잡한 절차의 수행, 단순한 절차의 수행, 문제해결, 의사소통과 추론의 5개 영역이 포함된다(서동엽, 2000). 수행 영역 중에서 지식 영역은 위의 교육성취도 평가의 행동 영역 중 이해 영역과 대등한 것으로 볼 수 있으며, 교육 성취도 평가의 계산 영역을 보다 세분하여 일단계 계산인지 아니면 다단계 계산인지의 여부에 따라서 각각 단순한 절차의 수행 영역과 복잡한 절차의 수행 영역으로 구분하고 있는 것으로 보인다. 그리고, 추론과 의사소통을 TIMSS-R에서는 하나로 묶어 두고 있다.

TIMSS-R에서 이용된 문항 중에는 두 가지 이상의 내용 영역을 통합한 성격의 문항이 상당 수 포함되어 있으며, 이 점 때문에 교육성취도 평가 연구에서 지적했던 내용 영역의 범위 문제를 동일하게 언급할 수 있다.⁷⁾ Mullis et al.(2000)에서는 통합적인 문항에 대하여 중심적인 아이디어가 포함된 영역을 기준으로 분류하여 각 내용 영역에 따른 결과를 제시하고 있다. 이 결과로부터 TIMSS-R에서는 통합적인 문항을 다수 포함하고 있기는 하지만, 그 결과 해석에서는 통합적인 문항을 별도로 중요하게 다루지 않았다는 결론을 얻을 수밖에 없다. 한편, TIMSS-R의 수행 영역을 교육성취도 평가 연구의 행동 영역과 비교하면, 계산 영역은 두 영역으로 세분되고, 추론 영역과 의사소통 영역은 한 영역으로 통합된 것으로 볼 수 있다. 하지만, Mullis et al.(2000)에서는 행동 영역과 관련된 어떠한 결과도 언급하고 있지 않기 때문에 수행 영역에 대한 정보를 얻는데는 한계가 있다.

한편, 학생들의 영재성 측정과 판별에 관한 연구를 수행한 송상현(1998)의 연구에서는 내용 영역과 사고 능력이라는 이원 분류를 이용하고 있다. 내용 영역은 수, 통계, 관계, 수계열, 도형으로 나뉘어져 있어 수계열 영역을 제외하고는 우리 나라의 제 6차 교육 과정의 내용 영역과 일치한다. 또한, 사고 능력은 정보의 조직화 능력, 비례 추론 능력, 연역적 추론 능력, 귀납적 추론 능력, 수학적 추상화 능력, 시각화/공간화 능력, 직관적 통찰력으로 나누고 있다.⁸⁾ 내용 영역이 교육과정의 내용 영역에 바탕하고 있다는 점은 다른 평가와 유사한 특징으로 보인다. 또한, 사고 능력으로 표시된 분류는 Krutetskii(1976)가 분류하고 있는 영재의 특징적인 사고 능력에 바탕하여 도출된 것으로, 교육성취도 평가 연구에서 제시하고 있는 행동 영역에 대응되는 것으로 생각해 볼 수 있다. 다만, 우정호(1998)가 제시하고 있는 4가지 수학적 능력은 일반적인 수학교육에서 목표로 하는 것임에 비하여, 송상현(1998)의 연구에 제시된 사고 능력은 영재의 특징에 국한된 것이어서 영역 구분에서 상당한 차이를 보이고 있는 것이다.

특히, Krutetskii(1976)는 신속하고 정확한 계산 능력은 영재의 특징과 관련이 적다고 밝히고 있어, 일반적인 수학적 사고 능력에 계산 능력이 포함되는 것과는 차이가 있다. 그리고, 정보의 조직화 능력이나 수학적 추상화 능력, 시각화/공간화 능력, 직관적 통찰력 등은 우정호(1998)가 제시한 수학적 사고 능력을 다른 차원에서 분류한 것으로 볼 수 있으며, 추론 능력을 보다 세분하여 연역적 추론, 귀납적 추론, 비례 추론으로 세분하고 있는 점도 차이가 있다. 한편, 송상현(1998)은 한 문항의 사고 능력을 나타내면서 연역적 추론과 직관적 통찰의 두 가지를 이용하고 있기도 하다. 이 점은 교육성취도 평가의 내용 영역에서 제

7) 박정 외(2000)에서는, TIMSS-R에 출제된 162개의 문항 중에서 우리 나라 교육 과정을 기준으로 하여 우리와 다른 관점을 취하거나 2개 이상의 내용 영역의 통합적인 문항이 모두 37개가 있었다고 밝히고 있다. TIMSS-R의 내용 영역이 아닌 우리 교육 과정의 내용 영역을 기준으로 분류했다는 점과 우리 교육 과정과 다른 관점을 취하는 문항도 포함되었다는 점은 감안하더라도, 적어도 30개 정도의 통합적인 문항이 출제되었다고 볼 수 있다.

8) 송상현(1998)의 연구에서 이원분류를 명시적으로 제시한 것은 아니며, 문항분류표로부터 이원분류에 해당하는 부분을 발췌한 것이다. 또한, 1부와 2부 검사로 나누어진 판별 도구 중 1부 검사만을 참고한 것이다. 2부 검사에서는 영재의 특징 중 유창성, 융통성, 독창성을 검사하고 있고, 이를 확대하여 사고 능력에 포함할 수도 있겠으나, 이원분류의 형태가 보다 명확한 1부 검사만을 참조한 것이다.

기되었던 통합적인 성격의 문항에 대하여 두 가지를 동시에 제시한 것으로 해석할 수 있다.

지금까지 살펴본 세 가지의 평가의 실체는 각각 국내 평가와 국제 평가, 그리고 특수한 아동에 대한 평가에 대한 사례에서 도출한 것이다. 세 가지 평가로부터 알 수 있는 공통점이라면 용어에서 다소 차이는 있지만 내용 영역과 행동 영역이라는 이원분류를 활용하고 있다는 것이 될 것이다. 그리고, 내용 영역에서는 대체로 교육과정의 내용 영역을 이용하고 있다는 것도 공통적이라고 볼 수 있으며, 교육성취도 평가 연구와 TIMSS-R에서 살펴본 것과 같이 일반적인 수학적 능력의 평가에 있어서는 계산, 이해, 추론, 문제해결, 의사소통이라는 교육성취도 평가의 행동 영역이 무난한 것으로 보인다.

여기서, 이러한 이원분류가 이용되고 있는 이유를 면밀히 생각해 볼 필요가 있다. 서두에서 밝힌 바와 같이 Tyler와 Bloom의 영향으로 이러한 이원분류가 시작되었고, 현재 이루어지고 있는 다양한 평가 장면에서 이원분류가 이용되고 있는 것이 실상이지만, 단지 전통이기에 지금까지도 쓰인다고 보는 것은 무리가 있다. NCTM에서는 평가에서 내용-행동의 이원분류 행렬에 의해 조직화된 특수하고 개별화된 많은 기능에 초점을 두기보다는 광범위한 범위의 수학적 과제에 초점을 두고 수학의 전체적인 관점을 택할 것을 권고하고 있기도 하다(구광조 외 공역, 1992). 이 권고의 초점은 단순한 기능을 평가하기보다는 통합적이고 전체적인 관점의 평가를 더욱 중요시하는 데 있는 것으로 보이며, 기능을 평가하는 수단으로서의 이원분류의 사용에 대한 관심은 감소되어야 할 것으로 보고 있다. 다만, 추론이나 문제해결과 같은 행동 영역에 대한 평가가 기능을 평가하는 것으로 보기는 어려우므로, 이러한 구분이 NCTM의 권고와 서로 대치되는 것으로 보기는 무리가 있다. 보다 중요한 것은 어떠한 종류의 평가이든 간에 그 평가를 통하여 평가자가 의도하는 목적이 있을 것인 바, 이 목적에 적절한 평가를 구현하기 위한 보조 수단으로서 도움이 될 때 이원분류는 진정으로 의미가 있을 것이라는 점이다. 이를 위하여 다음절에서는 그 동안 이루어진 평가의 실제에 대한 결과를 분석함으로써 원래의 평가의 목적을 구현하는 과정에서 이러한 이원분류가 어떻게 활용되고 있는지를 살펴보고자 한다.

2. 수학교육에서 평가의 목적과 결과 해석의 실제

교육성취도 평가의 목적은 네 가지로서, 국가 수준의 교육성취도 파악, 교육성취도의 추이 파악, 교육성취도와 배경 변인간의 관계 분석을 통한 교육 과정의 질 관리, 질 높은 양질의 문항 및 평가 기법의 개발로 요약될 수 있다(김명숙 외, 1999; 황혜정, 2000). 이 중에서 교육성취도의 추이 파악은 다년간의 교육성취도 평가 결과에 대한 종단적인 연구를 뜻하며, 교육성취도와 배경 변인간의 관계 분석은 수학 교육 이외의 분야와의 연구를 필요로 하므로 본 고에서는 교육성취도 파악과 문항 및 평가 기법의 개발이라는 목적에 국한하고자 한다. 교육성취도 평가는 이러한 목적 하에 앞의 절에서 살펴본 바와 같은 내용 영역과 행동 영역을 설정하여 평가를 시행하였고 그 결과를 분석하여 제시하고 있다.⁹⁾ 김명숙 외(1999)에서는 정답률을 포함한 문항별 통계 처리 결과를 제시하고 있으며, 황혜정(2000)에서는 내용 영역별로 전체 학생들의 학력이 우수학력, 보통학력, 기초학력, 기초학력미달 중에서 어디에 해당하는지를 프로파일 형태로 제시하고 있고, 대표적인 문항에 대하여 문항별 학생 반응과 전반적인 결과를 소개하고 있다. 이러한 형태의 결과 제시와 관련하여 두 가지의 문제를 논하고자 한다.

첫째는, 행동 영역에 대한 결과 제시가 미흡하다는 것이다. 김명숙 외(1999)에서 여러 연구에 기초하여 계산, 이해, 추론, 문제해결, 의사소통이라는 5개의 행동 영역을 설정하고 이에 기초한 문항 분포를 제시하

9) 황혜정(2000)에서는 문항에 대한 행동 영역은 제시하지 않고 있다. 그러나, 김명숙 외(1999)의 연구가 2000년의 교육성취도 평가를 위한 예비 문항을 개발하는 연구였음을 감안하면 동일한 행동 영역이 설정되어 있었을 것이다.

있지만, 이러한 행동 영역의 구분에 따른 결과는 명시적으로 드러나 있지 않다. 다만, 황혜정(2000)에서 '학생들은 정확한 답을 구하는 데에는 익숙하지만 그 이유나 결과를 말이나 글로 설명하는 데에는 어려움이 따르는 것으로 판단된다'고 하여, 교육성취도 평가에 포함되었던 서술형 문항에 대한 결과를 바탕으로 의사소통 영역에 대한 결과를 언급하고 있을 뿐이다.¹⁰⁾ 김명숙 외(1999)에서는 초등학교 및 중학교, 고등학교 각각 5개 행동 영역별 문항 분포가 어떠한지를 명시하고 있으나, 결과 분석에서 적극적으로 활용되지 않았다는 점은 이러한 행동 영역을 구분하는 의미를 재고해 보게 한다.

평가에서 행동 영역을 구분하는 의도는, 구분된 각각의 행동 영역에 대한 문항을 고르게 안배한다는 것과 더불어 행동 영역에 따른 결과를 제시할 수 있을 때 더욱 그 의미가 커지게 될 것이다. 그리하여, '학생들이 전반적으로 계산 능력에서는 우수하지만 추론 능력에서 우수하지 않다'는 식의 결론을 내릴 수 있고, 이 결론이 타당하게 받아들여질 때 의미가 있을 것이다. 특히, 교육성취도 평가의 행동 영역은 학생들의 수학적 사고력에 대한 모종의 해석을 기대한다는 점에서, 교육과정의 내용 영역별 성취도와 더불어 각각의 수학적 사고 영역에 대한 결과를 알아보는 것이 의미가 있을 것이기 때문이다.

둘째는, 내용 영역별 결과를 분석함으로써 우리가 어떤 해석을 내릴 수 있는지의 문제이다. 이를테면, 황혜정(2000)에서는 중학생들의 성취 수준에 대하여 도형 영역과 함수 영역은 기초학력인 것으로, 그 외의 4개 영역은 보통학력인 것으로 프로파일을 제시하고 있다. 이러한 성취수준 결정의 근거의 예를 들면, 도형 영역의 경우 14점 만점인 8개의 문항 중에서 평균 점수가 3~5점 사이에 있었다는 것이다. 교육성취도 평가에서는 이렇게 각각의 영역의 성취수준을 결정하는 근거를 제시하고 있으며 그 근거에 의하여 학생들의 학력을 판단하고 있으나, 문제는 우리가 이 결과로부터 일반적으로 우리 나라 학생들은 도형 영역에 대한 성취 수준이 보통에도 이르지 못하고 있다는 결론을 내릴 수 있는지는 점이다. 이러한 판단은 문항에 대한 검토와 함께 이루어져야 할 것이므로 다음 장에서 논의하기로 한다.

교육성취도 평가 연구에 대하여 본 연구에서 제기하는 문제는 TIMSS-R에서도 비슷한 형태로 반복되고 있다. TIMSS-R의 최종 결과 보고서인 Mullis et al.(2000)에 제시된 결과는 전체적인 성취도, TIMSS와의 추이 분석, 내용 영역별 성취도, 수준점별 성취도, 성별 차이, 교육과정, 학교제도, 배경 변인 등에 대한 국가간 비교로 이루어져 있다. 서론에서 밝힌 바와 같이 국가간 성취도와 교육과정을 비교하고 교육과정 및 배경 변인과 성취도 사이의 패턴을 찾는다는 목적에는 적절한 결과 보고서일 것이나, 내용 영역과 수행 영역과 결부시켜 생각하면 교육성취도 평가 연구와 비슷한 두 가지 문제를 제기할 수 있다.

첫째는, Mullis et al.(2000)에서 제시하는 결과에는 수행 영역에 따른 결과가 언급되어 있지 않다는 점이다. 1995년에 시행되었던 TIMSS에서는 각 문항에 대한 수행 영역을 명시하였음에 비하여 1999년에 시행된 TIMSS-R에서는 수행 영역별 문항의 총 개수만 제시하는 등의 변화가 있기는 하였지만, 평가 시행 당시에는 있었던 수행 영역이 결과 분석에서 배제된 것이다. 이는 교육성취도 평가에서와 마찬가지로 수행 영역을 구분하는 의미를 재고해 보게 한다.

둘째는, 내용 영역별로 국가별 순위를 제시한 결과에 우리가 어떤 의미를 부여할 수 있는지는 것이다. TIMSS-R의 결과, 우리나라의 전체 성취도는 싱가포르에 이어 2위였으며, 자료의 표현 및 해석·확률 영역에서는 1위를, 분수와 수 감각 영역에서는 4위를, 나머지 5개 영역에서는 2위인 것으로 기록되어 있다. 하지만, 두 가지 점 때문에 이 순위로부터 우리나라 학생들의 수학적 능력이 2위라는 해석을 하기는 어렵다. 우선, 1절에서 지적한 바와 같이 TIMSS-R에는 상당수의 통합 영역 문항이 포함되어 있었기 때

10) 교육성취도 평가 연구는 전국적인 시험이고 그 결과가 일선 학교와 학생들에게 주는 파급력이 크기 때문에 자세한 결과는 일반에 공개하지 않는 것이 원칙이다. 하지만, 행동 영역에 따른 결과 분석이 명시적으로 이루어진 것이 없다는 점은 분명해 보인다.

문에 TIMSS-R에서 구분한 분수와 수 감각 영역에 대한 우리 나라의 순위가 2위였다는 해석은 가능하지만, 이로부터 일반적인 분수와 수 감각 영역에 대한 우리 나라 학생들의 능력이 2위라는 해석은 확실하지가 않다는 것이다. 또한, 각각의 영역에 대하여 TIMSS-R에 참가한 학생들이 학습한 내용의 범위나 접근 방법에 차이가 있었기 때문에 위와 같은 해석은 곤란하다는 것이다.

NCTM에서 제시하는 수학교육 평가에 대한 권고 사항 중의 하나는 ‘평가를 통하여 단지 등급을 정할 목적으로 옳은 답의 개수를 세기보다는 평가를 가르치는 것에 통합시켜야 한다’는 것이다(구광조 외 공역, 1992). 어떤 문항을 이용하든 어떤 평가의 결과로 순위를 부여하는 것은 가능하다. 하지만, 중요한 문제는 이 결과로부터 우리 나라의 수학 수업에 어떤 시사점을 얻을 수 있느냐는 것이 될 것이다. 이를 위해서는 우리 나라 학생들의 성취도가 상대적으로 높은 문항과 낮은 문항을 분류하여, 높은 문항이 높게 나온 이유는 무엇이며 낮은 문항이 낮게 나온 이유는 무엇인지, 그리고 우리는 낮지만 높게 나온 국가에서 채택하는 방법은 어떠한 것인지를 조사하는 일이 필요해지게 될 것이다.

한편, 교육성취도 평가 연구나 TIMSS-R과는 대조적으로 송상현(1998)의 연구에서는 영재아의 특성을 나타내는 사고 능력을 기준으로 상세한 결과를 제시하고 있다. 이는 영재아의 특성에 대한 여러 연구에서 제시하는 결과가 주로 사고 능력에 관한 것이기 때문이었을 것으로 보인다. 또한, 그렇기에 어떤 학생이 영재인지를 판단하는 데 있어서, 예를 들어 정보의 조직화 능력은 상위 1% 정도의 수준이지만, 직관적 통찰력은 상위 10% 정도의 수준이라는 해석이 의미 있게 되는 것이다. 이 연구에서 이용하였던 내용 영역에 따른 결과 제시가 없기는 하지만, 이는 문항의 설명을 위한 도구로서만 활용된 것으로 보이며, 영재의 특성과는 큰 상관이 없다는 점에서 문제는 없어 보인다.

세 가지의 평가에 대한 고찰로부터, 평가의 목적을 반영하는 이원 분류를 설정하고 이에 따라 평가의 결과를 분석하며, 분류를 통한 평가 결과를 의미 있게 하는 일이 현재까지는 그리 잘 이루어지지 않고 있음을 알 수 있다. 그 형태가 일원분류이든 아니면 이원분류, 삼원분류이든 간에 중요한 것은 평가의 목적에 대한 결과를 의미 있게 도출하는 데 적절히 활용될 때 이러한 분류가 유용하다는 것이다. 이를 위하여 평가를 시행하기 이전에 평가의 목적을 결정하는 과정에서 어떤 형태의 분류를 이용할지를 면밀한 검토 후에 결정할 필요가 있다. 예를 들어, 제 7차 교육과정의 평가 항목에서는 다음과 같은 내용을 명시하고 있다(교육부, 1997a).

아. 수학과 학습에서 전반적으로 요구되는 다음 사항을 강조하여 평가한다.

- (1) 수학의 기본적인 개념, 원리, 법칙의 이해.
- (2) 수학의 용어와 기호를 정확하게 사용하고 표현하는 기능.
- (3) 수학적 지식과 기능을 활용하여 문제를 수학적으로 사고하여 해결하는 능력.
- (4) 실생활 현상을 수학적으로 관찰, 분석, 조직, 사고하는 태도.

따라서, 제 7차 교육과정에 따라 차상급 단계로의 진급 여부를 결정하는 평가라면, 내용 영역에서는 교육 과정의 내용 영역과 실제 지도 시간을 고려하여 문항 수의 비율을 결정하고, 위의 네 가지 사항을 참조하여 행동 영역을 결정할 수 있을 것이다. (1)번과 (2)번 항목은 각각 교육성취도 평가에서 ‘이해’와 ‘의사소통’이라고 이름 붙인 영역과 일치하는 것으로 보인다. (4)번 항목도 ‘외적 문제해결’이라고 이름을 붙인다면 무난할 것으로 보인다. 하지만, (3)번 항목에 대해서는 대략적으로 ‘수학적 사고력’이라고 이름 붙일 수 있을 것이나, 구체적인 항목에 대한 추가적인 논의가 필요할 것으로 보인다. 앞서 언급한 우정호(1998)의 구분으로는 수학적 사고에 계산, 이해, 추론, 문제해결이 모두 포함되나, 이해와 외적 문제해결은

각각 (1)번과 (4)번 항목에서 구분된 것이다. 결과적으로, 계산과 추론, 내적 문제해결이 (3)번 항목에 남게 된다. 결과적으로, 계산, 이해, 추론, 의사소통, 내적 문제해결, 외적 문제해결의 6개 영역으로 구분한다면 교육과정에 제시된 내용을 반영할 수 있을 것으로 보인다. 이와 더불어, 각각의 영역에 해당하는 문항의 비율은 해당 단계의 교육과정과 교과서를 참조하여 결정하는 것이 무난하리라 생각된다. 아울러, 각각의 영역에 대한 더 이상의 세분은 교육과정과 교과서를 참조하는 것이 바람직할 것이다. 예를 들어, 추론 영역의 경우 초등학생 대상의 평가는 귀납 추론이 중심이 되어야 할 것이며, 고등학생의 평가라면 교육과정에서 다루는 비중에 따라서 연역 추론과 귀납 추론을 활용할 수 있다는 것이다.

하지만, 이렇게 하더라도 내용 영역이나 행동 영역에서 통합적인 성격의 문항을 다루는 문제가 남게 된다. 그것은 교육과정과 교과서에서는 세부 내용 영역과 단원별로 내용이 구성되기 때문에 통합적인 내용을 다루기 어렵게 되어 있기 때문이다. 그래서, 평가자가 평가의 목적을 고려하여 통합적인 문항을 어느 정도로 다루는 것이 적절한 지를 결정하는 것이 바람직할 것으로 보이며, 이러한 기준을 마련하기 위해 많은 연구가 필요하다고 생각된다. 또한, 이 경우에는 통합적인 문항을 별도의 영역으로 구분하는 것이 바람직할 것이다.

III. 지필 평가 문항의 실제

김재춘 외(2000)는 평가 문항을 제작하는 과정에서 일반적으로 고려할 사항으로서 다음의 여섯 가지를 들고 있다.

- 첫째, 문항을 풀기 위해 필요한 능력이 검사 계획서 상의 내용과 행동에 부합되는가?
- 둘째, 문항의 내용이 모호하거나, 부적절한 용어가 있거나, 어색한 문장이 있지 않은가?
- 셋째, 문항이 필요 이상으로 장황하지 않은가?
- 넷째, 여러 전문가들이 정답에 정답으로 동의할 것인가?
- 다섯째, 일반적 문항 제작의 원리에 벗어나지 않는가?
- 여섯째, 문항이 특정 학생들에게 편파적이지 않은가?

이상의 여섯 가지 요소는 평가 문항이 갖추어야 할 기본 요건으로 볼 수 있을 것이며, 본 고에서 살펴 보고 있는 교육성취도 평가나 TIMSS-R 등의 평가에 이용된 문항은 이러한 기본 요건을 충실히 갖춘 문항으로 간주할 수 있을 것이다. 이 장에서 다루고자 하는 문제는 이러한 기본 요건을 갖춘 문항을 이용하여 평가한 결과를 해석하는 것과 관련되며, 여러 지필 평가에 이용된 문항을 분석함으로써 평가자가 의도했던 목적에 맞추어 결과를 해석하는 데 있어 문제가 될 수 있는 요소를 분석해 보고자 하는 것이다. 이를 위하여 첫째로, 교육과정을 기준으로 한 성취 정도를 파악하고자 하는 평가에 이용된 문항을 분석하고, 둘째로, 국제 비교 평가에서 이용된 문항을 분석할 것이다.

1. 성취 정도를 파악하는 평가 문항의 실제

제 7차 교육과정에서 2-가 단계의 수와 연산 영역에서 ‘덧셈과 뺄셈의 활용’에 대하여 서술된 내용은

‘덧셈과 뺄셈을 활용하여 여러 가지 실생활 문제를 해결할 수 있다’는 것이다(교육부, 1997a). 2학년 학생이 이 기준에 대하여 도달하였는지 여부를 지필 평가 문항을 이용하여 평가한다고 하자. 제 7차 교육과정의 평가 항목에서는 상, 중, 하의 평가 기준을 명시하고 있으며, 이를 반영하여 제작된 평가 기준의 예가 김정호 외(1999)에 다음과 같이 제시되어 있다.

상 : 두 자리 수의 덧셈, 뺄셈, 혼합 계산이 적용되는 문제를 능숙하게 해결할 수 있다.

중 : 받아 올림이 있는 두 자리 수의 덧셈이 적용되는 문제를 해결할 수 있다. 받아 내림이 있는 두 자리 수의 뺄셈이 적용되는 문제를 해결할 수 있다.

하 : 받아 올림이 있는 두 자리 수의 덧셈을 계산할 수 있다. 받아 내림이 있는 두 자리 수의 뺄셈을 계산할 수 있다.

위의 평가기준은 교육과정에 서술된 내용을 충분히 반영한 것으로 보인다. 그러나, 중 수준에 도달하였는지 여부를 위한 평가 문항을 제작한다고 할 때 어떻게 해야 할 것인가? 다음의 네 문항을 비교해보자.

(가) 약수터에는 어른이 36명, 어린이가 17명 있습니다. 약수터에는 모두 몇 명이 있습니까?(교육부, 2000 : 57).

(나) 약수터에는 남자 어린이가 36명, 여자 어린이가 17명 있습니다. 약수터에는 모두 몇 명의 어린이가 있습니까?

(다) 약수터에서 남자 어린이 36명이 놀이를 하고 있었습니다. 얼마 후에 남자 어린이 17명이 더 왔습니다. 약수터에 있는 남자 어린이는 모두 몇 명입니까?

(라) 약수터에서 어린이 36명이 놀이를 하고 있었습니다. 얼마 후에 남자 어린이 17명과 여자 어린이 18명이 더 왔습니다. 약수터에 있는 어린이는 모두 몇 명입니까?

(가)의 문항을 해결하기 위해서는 $36 + 17 = 53$ 의 계산을 수행해야 하는 것 외에도 ‘모두 몇 명’이라는 말에서 약수터에 있는 ‘모든 사람의 수’를 계산해야 한다는 것, 그리고 사람은 어른과 어린이로 분류될 수 있다는 것을 알아야 한다. 전체와 부분의 관계에 대한 Piaget의 실험 중에서 전조작기의 아동들은 가역성의 부재로 인하여 ‘앵초 꽃 7송이와 다른 꽃 3송이가 있을 때, 앵초 꽃이 많은가, 꽃이 많은가’라는 문제를 해결하지 못한다는 것을 보여 주는 결과가 있다(김응태 외, 1984). 교육과정의 문구에 제시된 ‘실생활 문제’라는 표현에는 이러한 분류 상황도 포함될 것으로 보이지만, 어느 정도까지의 분류를 요구하는지는 명시되어 있지 않으며, 명시하기도 어려운 내용으로 보인다. 그렇기에 사람은 어른과 어린이로 나누어진다는 것을 묻기보다는 (나) 또는 (다)의 문항과 같이 보다 더 명시적인 분류를 요구하는 것이 보다 바람직할 것으로 생각된다.

그러나, (나)와 (다)의 문항에서도 차이는 있을 수 있다. 강지형 외(1999)는 이러한 두 상황을 ‘합병’ 상황과 ‘첨가’ 상황으로 구분하고 있으며, 각각 합집합의 개념과 피아노 공리의 귀납적인 개념을 요구하는 것으로 분석하고 있다. 그렇기에 학생들이 두 문항을 해결하기 위하여 덧셈 계산이 필요함을 인식하기 위해 요구되는 사고는 차이가 있을 수 있다는 것이다. 2-가 단계의 교과서에서는 두 가지 상황 모두를 다루고 있으며, 그렇기에 가능하다면 두 유형의 문항을 모두 포함하는 것이 바람직할 것으로 생각된다. 그러나, 두 문항 중에서 어느 한 문항만을 해결한 학생에 대하여 교육과정의 기준에 도달하였다고 판단할 수 있는지의 문제는 여전히 어려운 문제로 남게 된다.

한편, (라)의 문항을 해결하기 위해서는 $36 + 17 = 53$ 이라는 계산을 수행한 다음 $53 + 18 = 71$ 의 계산을 한 번 더 수행하여야 한다. 즉, (가)의 문항 A는 일단계 문제인 반면 (나)의 문항은 다단계 문제이며 (강완 외, 1998), 문제를 해결하기 위해 필요한 연산 능력은 동일하다고 하더라도 학생들의 성취도에 차이가 생길 것임은 분명해 보인다. 여기서 지적하고자 하는 문제는 (가)의 문항을 해결하였지만 (라)의 문항을 해결하지 못한 학생이 있을 때, 교육과정의 기준에 도달했다거나 도달하지 못했다는 판단을 내릴 수 있는지는 것이다. 교육과정에 제시된 문구에서는 덧셈과 뺄셈 연산을 몇 번까지 적용해야 하는지는 언급하고 있지 않다. 그렇다고 하여 이까지 언급하는 것이 바람직한 것인지도 논의할 문제가 된다.

다음으로 들고자 하는 예는 6-가 단계의 규칙성과 함수 영역 중 '비례식'에 대한 내용이다. 비례식에 대하여 교육과정에 서술된 내용은 '비례식을 이해하고, 이를 활용할 수 있다'는 것이다. 이 중에서 비례식을 활용하는 것과 관련하여, TIMSS-R에 출제되었던 문항 중에서 두 문항과 그 결과를 비교해 보자(김성숙 외, 1999).¹¹⁾

A4. 연주가 운동장을 4바퀴 도는 동안 영희는 3바퀴를 돈다. 영희가 운동장 12바퀴를 돌면 연주는 몇 바퀴를 돌겠는가?

- ① 9 ② 11 ③ 13 ④ 16

(정답률 : 53.2%, ①을 선택한 학생의 비율 : 25.2%)

M6. 버스 안에 36명의 승객이 있다. 이 중 어린이와 어른의 비는 5 : 4이다. 버스에 타고 있는 어린이의 수를 구하여라.

(정답률 : 63.0%)

위의 A4 문항은 비례식의 활용에 대한 제 6차 교육과정기의 교과서에도 제시되어 있는 전형적인 문항이므로(교육부, 1997b), 비례식의 활용에 대한 학생들의 성취 정도를 파악할 때 전형적으로 이용할 수 있는 문항이다. M6 문항을 해결하기 위해서는 초등학교 6학년 학생의 수준에서는 표만들기 전략을, 중학교 1학년 학생은 일차 방정식을, 중학교 2학년 학생은 연립방정식을 이용할 수 있다. 그렇기에 M6 문항은 A4 문항보다는 '더 어려운' 문항이라고 볼 수 있다. 여기서, M6 문항의 정답률과 비교하여 A4 문항의 정답률이 낮은 이유를 A4 문항에서 보기 ①을 선택한 학생의 비율과 관련지어 보고자 한다. 보기 ①의 답을 얻은 학생들이 범할 수 있는 대표적인 오류는 문항에 제시된 숫자를 순서대로 이용하여 $4 : 3 = 12 : \square$ 라는 비례식을 세운 후에 답을 구한 것이다.¹²⁾ 아울러, 자유반응형 문항인 M6 문항의 정답률이 63.0%라는 사실은 비례식을 활용할 수 있으면서도 A4에서 보기 ①을 선택하여 오답을 한 학생이 상당수 있었을 것이라는 생각을 가능하게 해 준다.

여기서 제시된 정답률은 중학교 2학년을 대상으로 한 결과이므로, 위의 결과에 대하여 내릴 수 있는 해석은 적어도 중학교 2학년에 대해서는 비례식에서 제시되는 숫자의 순서를 바꾸는 것이 일차방정식이나

11) TIMSS-R의 평가 대상은 중학교 2학년이었으나, 우리나라의 제 6차 교육과정을 기준으로 하였을 때 초등학교 수준으로 볼 수 있는 문항이 67%가 포함되어 있었다(김성숙 외, 1999). 이것이 우리나라 학생들이 두 번째로 높은 성취도를 보이는 데 긍정적으로 작용하였는지는 단언하기 힘들다. 그것은 다른 국가에서는 초등학교 수준의 문항이 어느 정도였는지 파악하기 힘들고, 한편으로는 학습한 후에 오랜 시간이 지난 내용을 묻는 것이 과연 더 쉬운 쪽으로만 영향을 주는지가 명확하지 않기 때문이다.

12) A4 문항은 선다형 문항이므로 보기 ①을 선택한 학생 모두가 이런 오류를 범했다고 말하기는 무리가 있지만, 상당수의 학생이 이러한 오류를 범했을 것으로 보인다. 또한, 같은 이유로, A4 문항에 대하여 정확히 계산을 수행하여 정답을 맞춘 학생의 비율은 더 작을 수도 있다.

연립방정식을 활용하는 것보다 더 문항을 어렵게 하는 요인이 된다는 것이다. 제 6차 교육과정기의 교과서에서 다루는 전형적인 문항은 다음과 같은 것이다(교육부, 1997b : 55).

연필 4자루를 700원에 샀다. 이 연필 14자루를 사려면 돈이 얼마 필요한가? 또 3500원을 가지면 이 연필은 몇 자루 살 수 있는가?

이 문항은 비례식을 세우는 데 필요한 숫자를 자연스럽게 주는 문항과 거꾸로 주는 문항의 2가지를 포함하고 있다. 이렇듯 교과서에서는 두 가지 상황 모두를 다루고 있는 것이다. 그러나, 위의 결과는 초등학교 6학년 학생들에 대해서도, 숫자를 제시하는 순서가 학생들의 정답률에 상당한 영향을 주리라는 것을 보여 주고 있다. 따라서, 비례식의 활용 문항에서 숫자를 두 가지 순서로 제시하여 한 문항만 맞춘 학생이 교육과정의 기준을 성취했다고 볼 수 있느냐의 문제가 어려운 문제가 되는 것이다.

지금까지의 논의와 더불어 고려할 문제는, 교육과정에서 제시하는 내용의 기본 수준과 학생들의 실제 수준이 차이가 심한 경우의 평가 방법이다. 예를 들어, 8-나 단계의 도형 영역에서 삼각형과 사각형의 성질에 대하여 교육과정에서 서술된 내용은 '삼각형의 합동조건을 이용하여 삼각형과 사각형에 관한 간단한 성질을 증명할 수 있다'는 것이다(교육부, 1997a : 74). 이러한 진술에 근거하여 학생이 교육과정의 기준을 성취하였는지 판단하기 위한 평가를 시행한다면, '증명'을 요구하지 않는 문항을 이용하기는 매우 어려워 보인다. 그러나, 우정호(1994)나 서동업(1999)의 연구 결과에서 보여 주고 있는 것은 대부분의 교과서에 제시되어 있는 매우 기본적인 명제를 증명할 수 있는 중학교 2학년 학생의 비율이 30%를 넘지 않는다는 것이며, 두 연구에 근거한다면 적어도 8-나 단계의 도형 영역에 대한 성취 수준을 충족시키는 학생의 비율은 30%를 넘지 않을 가능성이 많은 것이다.

이는 구체적으로 제 7차 교육과정을 시행하면서 8-나 단계로부터 9-가 단계로의 진급을 결정하는 평가를 시행할 때 문제를 발생시킬 수 있다. 이 평가에서 증명을 할 수 있는 학생을 진급시키고, 그렇지 못한 학생은 재이수 또는 보충 학습을 제공하기로 한다면 평균적으로 70% 정도의 학생들이 보충 학습에 참여해야 한다. 또한, 보충 학습의 결과로 학생들이 증명을 할 수 있게 된다는 보장도 할 수 없을 것이다. 그렇다고 하여, 교육 과정에서 '증명할 수 있는' 능력을 요구하고 있는 진술을 무시하고 직관적으로 성질을 파악하는 정도이면 기준에 도달한 것으로 본다는 식의 결론을 내리는 것도 바람직하지 않아 보인다. 이렇듯 학생들이 전형적으로 어려워하는 단원의 내용에 대하여 어떤 식으로 성취 정도를 파악할 것인지도 중요한 문제로 부각될 수 있다는 것이다.

2. 국제 비교 평가의 실제

TIMSS-R과 같은 국제 비교 평가에는 세계의 여러 국가가 참여한다는 점에서 몇 가지 제한점을 지닌다. 교육과정의 차이로 인한 내용 영역의 문제는 2장에서 논의하였으므로, 이 절에서는 문항 번역 상의 문제와 관련된 실례를 분석하고자 한다. Hanna(1993 : 247)는 문항의 번역에서 생길 수 있는 차이를 다음의 여섯 가지로 분류하고 있다.

- ① 언어의 곤란도의 차이.
- ② 정확한 번역에 따른 수학적 용어가 존재하지만, 학생들의 교과서에서 이용되는 용어와는 다른 경우.
- ③ 어법에 따른 추상화 수준의 차이.
- ④ 의도하고 있는 의미의 차이.

- ⑤ 표기에서 강조점이나 명확성의 차이.
- ⑥ 번역 상의 사소한 오류(예컨대, 인쇄상의 오류는 의미에 영향을 주지는 않는다.)

이러한 번역 상의 문제와 관련하여 살펴 볼 문항은 2개이다.

(마) John sold 60 magazines and Mark sold 80 magazines. The magazines were all sold for the same price. The total amount of money received for the magazines was \$700. How much money did Mark receive?(Mullis et al, 2000 : 73)

이 문항은 물건의 가격과 관련된 영어 문항의 번역에 있어서의 어려움을 단적으로 보여 주는 예이다. 위의 문항에서 잡지 1권의 가격은 5 달러임을 알 수 있고, 1권에 5달러인 잡지나 달러 화폐 단위 자체는 영어권 국가에서는 매우 친숙한 상황이다. 그러나, 우리말로 번역하면서 몇 가지 어려움이 유발되게 된다. 위의 문항을 직역한다면 다음과 같은 번역이 가능하다.

영희는 잡지 60권을 팔고, 철수는 잡지 80권을 판매하였다. 잡지의 가격은 똑같다고 한다. 전체 판매 금액은 700 달러였다. 철수는 얼마나 벌었는가?

그러나, 위와 같이 번역하게 되면, 영어권 학생들에게는 ‘달러’라는 화폐 단위가 일상 생활에서 이용하는 단위인 반면 우리 나라 학생들은 그렇지 못하다는 점에서 공정하지 못하게 된다. 하지만, ‘달러’를 우리의 화폐 단위인 ‘원’으로 바꾸면 새로운 문제가 생겨나게 된다.

영희는 잡지 60권을 팔고, 철수는 잡지 80권을 판매하였다. 잡지의 가격은 똑같다고 한다. 전체 판매 금액은 700원이었다. 철수는 얼마나 벌었는가?

이 문항에서 잡지 1권의 가격은 5원이 되고 이는 현실과는 동떨어진 상황이 되는 것이다. 그렇다고 하여, 잡지 대신 다른 물건을 이용하고자 하더라도 5원 짜리 물건을 찾을 방법이 없으며, 현실성 있게 ‘700원’을 환율에 맞추어 ‘910000원’으로 바꾼다면, 원래 문제에 비하여 훨씬 큰 수의 계산을 요구하는 더 어려운 문항이 된다.¹³⁾ 이 문항에 대한 우리 나라 학생들의 평균 정답률은 69%로서 38개 참가국 중에서 4번째로 높았지만, 영어권 국가인 싱가포르보다는 15%가 낮았다(Mullis et al., 2000 : 73).

다음은 번역에서 (마)와는 다른 종류의 문제를 보여 주는 문항이다.

(바) There are 68 rows of cars in a parking lot. Each row has 92 cars. Which of these would give the closest estimate of the total number of cars in the parking lot?(Mullis et al., 2000 : 80)

이 문항에서 지적하고자 하는 것은 교육과정에서 이용하는 용어의 차이이다. 문제가 되는 구절은 ‘closest estimate’라는 부분이다. 이 용어의 직접적인 번역은 ‘가장 가까운 어림값’ 또는 ‘최선의 어림값’ 정도가 되겠지만, 우리 나라 교육과정에서는 이용하지 않는 용어이다. 어림과 관련이 있는 부분으로 우리나라에서는 초등학교 4학년에서 반올림과 올림, 버림을 다루며, 중학교 2학년에서 근사값의 사칙 계산을 다

13) 1달러를 1300원으로 계산한 것이다.

룬다(교육부, 1996b : 104-107, 김연식 외, 1998 : 25-27).¹⁴⁾ 반올림이나 어림, 버림이라는 용어 속에는 이미 ‘가장 가까운’ 근사값이라는 의미가 포함되어 있는 것이며, 이를 명시적으로 이용하는 경우는 없는 것이다. 다만, 어림이라는 용어는 길이를 처음 지도할 때 ‘어림’하여 ‘약 몇 cm’로 나타내는 내용을 초등학교 3학년에서 다루고 있다(교육부, 1996a : 91). 또한, 중학교에서 다루는 근사값의 계산은 위의 문항에서 요구하는 어림 계산과는 그 맥락이 다르다. 이 문항에 대한 우리 나라 중학교 2학년 학생들의 정답률은 82%로 38개국 중 5번째였으며, 가장 높았던 싱가포르의 94%와는 12%의 차이를 보이고 있다. 어림과 관련된 생김 문제는 김성숙 외(1999 : 110-112)에서도 찾아볼 수 있다.

U1. 미현이는 비이커의 물이 95℃에서 70℃까지 식는 데 걸리는 시간을 알아보기 위하여, 물의 온도가 5℃ 식을 때마다 걸린 시간을 측정하였다.

기록 간격	냉각 시간
95℃ - 90℃	2분 10초
90℃ - 85℃	3분 19초
85℃ - 80℃	4분 48초
80℃ - 75℃	6분 55초
75℃ - 70℃	9분 43초

비이커에 있는 물이 95℃에서 70℃까지 식는 데 걸린 전체 시간을 분으로 어림 계산하고 그 풀이 과정을 써 보아라.

이 문항에 대하여 각각의 시간을 분으로 반올림하여 $2 + 3 + 5 + 7 + 10 = 27$ (분)이라고 답한 우리 나라 학생은 8.8%로서, 모두 더하여 26분 55초를 구한 다음 반올림하여 27분을 구한 학생의 비율인 13.9%나 모두 더하여 26분 55초라고 써서 오답을 한 학생의 비율인 17.3%에 비하여 훨씬 낮았다(김성숙 외, 1999 : 111). 이는 어림과 관련된 문항에서 번역상의 어려움을 보여 주는 것으로 볼 수 있다.

그 외에 번역 상의 문제와 관련하여 마지막으로 지적하고 싶은 단어는 ‘fraction’이다. 영어 단어 ‘fraction’은 우리말로 번역될 때 상황에 따라서 ‘조각’, ‘부분’, ‘분수’, ‘비율’ 등으로 다양하게 번역될 수 있다. 분수는 등분할의 의미, 양의 의미, 몫의 의미, 비의 의미, 연산자의 의미로 다양한 측면에서 해석이 가능하며(강지형 외, 1999 : 174), 우리 나라에서 비율 지도의 맥락은 비의 값으로부터 도입하여 백분율이나 할푼리 등 소수 표현으로 고쳐 해석하는 데 중점을 두고 있다(교육부, 1996c : 94-101). 이로 인하여 학생들이 비율을 학습한 후에 비율에 대하여 갖는 직관적 표상은 분수보다는 소수이거나 백분율의 값일 가능성이 많아 보인다. TIMSS-R에 출제된 문항 중에는 ‘fraction’을 ‘비율’로 번역할 수밖에 없지만 답을 분수로 요구하는 것이 있었는데, 예를 들면, 김성숙 외(1999 : 400)에 제시된 K1 문항 같은 경우이다. 문항은 ‘다음 원에서 어두운 부분이 차지하는 비율이 위의 직사각형에서 어두운 부분이 차지하는 비율과 대체로 같은 것은?’이며, 주어진 직사각형에서 나오는 $\frac{5}{12}$ 라는 비율에 가까운 그림을 찾아야 답할 수 있는 문항이었다. 이를 영어권의 학생들은 ‘fraction’이라는 단어에서 자연스럽게 분수로 사고할 수 있는 반면, 우리나라 학생들은 ‘비율’이라는 단어로 인하여 혼란을 겪었을 가능성이 있는 것이다. 그렇다고 해서 위의 문항을 ‘비율’ 대신 ‘분수’라는 말을 이용해서 서술하는 것은 매우 어색한 표현이 된다.

14) 국내에서 TIMSS-R가 시행된 1999년 2월의 중학교 2학년 학생들은 초등학교의 대부분과 중학교 과정을 6차 교육 과정에 따라 학습하였기 때문에, 6차 교육과정에 따라 구성된 교과서를 참조하였다.

따라서, 국제 비교 평가에 참여할 때 번역으로 인하여 원래 평가 문항의 곤란도에 주는 영향을 최소화할 수 있도록 하는 일이 평가 본연의 취지를 살리기 위하여 매우 중요한 문제가 된다고 하겠다.

IV. 결론 및 제언

본 연구에서는 수학교육에서 이루어지고 있는 여러 지필 평가 중에서 주로 교육성취도 평가 연구와 TIMSS-R을 소재로 하여, 평가 틀과 결과 해석, 평가에 이용된 문항을 중심으로 평가의 실제를 분석해 보았다. 본 고에서 이루어진 분석의 결과는 다음과 같이 요약될 수 있다.

첫째, 여러 지필 평가에서 내용 영역과 행동 영역 또는 이와 대등한 형태의 이원 분류가 이용되고 있지만 그리 적절하게 이용되지 못하고 있는 것으로 보인다. 우선, 내용 영역은 주로 교육과정에서 이용되는 내용 영역의 구분을 이용하고 있으며 이것이 바람직해 보이지만, 최근 들어 강조되고 있는 통합적인 문항을 평가에 활용하고자 할 때 보다 적절히 분류하기 위한 방법을 모색할 필요가 있는 것으로 보인다. 그리고, 행동 영역은 대체로 계산, 이해, 추론, 문제해결, 의사소통이라는 분류가 이용되고 있으나 내용 영역의 문제와 유사하게 어떤 문항은 두 가지 이상의 행동 영역에 포함될 수 있다는 것이 문제점으로 보인다. 그렇기에, 먼저 행동 영역의 구분이 서로 배타적으로 되어야 하는지에 대한 연구가 필요할 것이며, 그렇지 않다면 통합적인 성격의 문항을 분류하기 위한 방법을 모색할 필요가 있을 것이다. 아울러, 교육과정을 기준으로 학생들의 성취 정도를 파악하는 평가에서는 무엇보다도 교육과정과 교과서를 바탕으로 이러한 영역을 설정할 필요가 있을 것이다.

둘째, 여러 지필 평가의 결과를 해석하는 데 있어서 나타나는 특징의 하나는, 문항을 제작하는 단계에서 활용되는 행동 영역에 대한 결과의 해석이 미흡하다는 것이다. 평가의 목적에 따른 결과 해석 과정에서 의미 있는 결론을 도출할 수 있도록 행동 영역을 설정하는 방법을 찾을 필요가 있을 것이며, 평가의 성격에 따라서는 행동 영역을 이용하지 않는 일원 분류를 이용하거나 다른 차원까지 고려하는 삼원 분류를 활용할 수도 있을 것이다. 문항을 제작하는 과정에서 문항의 수를 고르게 분포시킬 목적으로만 행동 영역을 활용하는 것은 바람직하지 않은 것으로 보인다. 또한, 내용 영역에 따른 결과를 제시할 때에도 통합적인 성격의 문항에 대한 처리는 여전히 문제가 되며, 내용 영역을 반영하여 출제된 문항이 어떠한지 지대한 영향을 미칠 수 있다는 점을 감안해야 한다.

셋째, 교육과정을 기준으로 학생들의 성취 정도를 파악하는 평가에서 이용되는 문항에 대한 세부적인 기준을 심사숙고하여 결정할 필요가 있을 것이다. 교육과정에 서술된 내용만을 기준으로 한다면 구체적인 문항의 형태는 매우 다양해질 수 있으며, 이로 인하여 학생들의 정답률에서도 상당한 차이가 유발될 수 있음을 살펴보았다. 저학년의 덧셈 활용 문항에서 이용되는 소재의 분류 관계나, 문항에 대한 전형적인 풀이 과정을 고려하여 조건을 자연스러운 순서나 역 순서로 제시하는 것에 따라서 학생들의 반응은 상당히 차이가 있을 수 있다는 것이다. 그렇기에 성취 정도를 판단하는 평가에서는 보다 구체적인 성취 판단 기준을 설정하는 것이 필요할 것이다. 또한, 교육과정에 기술된 내용의 수준에 도달하지 못하는 학생이 상당수 존재하는 것으로 이미 밝혀진 내용에 대한 평가를 시행하는 경우, 어느 정도의 기준을 적용하는 것이 바람직한지를 결정하기 위한 연구도 필요할 것으로 보인다.

넷째, 국제 비교 평가에서는 주로 번역의 어려움으로 인한 문제가 유발될 수 있음을 살펴보았다. 문항에 화폐 단위가 이용되는 경우 영어권 문화와의 차이로 인하여 정답률에서 큰 차이가 유발될 수 있으며, 교육과정에서 유사한 내용을 다루더라도 접근 방법의 차이나 주로 강조하는 내용의 차이로 인하여 평가 결

과에서 차이를 보일 수 있다는 것이다. 국제 비교 평가에서 보다 공정한 비교를 위해서는 이러한 차이를 극소화할 수 있도록 문항 제작 과정에서 노력을 들일 필요가 있는 것으로 보인다.

수학교육이 이루어지는 다양한 장면에서 평가가 활용되고 있다. 그러나, 본 연구에서 살펴보았던 여러 평가의 실재를 분석해 본 결과, 평가 목적을 구현하는 데 있어서 평가 틀을 적절히 설정하고 적절한 수준의 문항을 출제하는 데 많은 어려움이 있음을 알 수 있었다. 평가 틀과 문항에 대한 합리적이고 구체적인 기준을 마련하는 일이 평가에서 시급한 과제라고 생각되며, 이 작업이 이루어졌을 때 원래의 평가 목적에 기초한 결과의 해석이 보다 의미 있게 될 것이다.

참고 문헌

- 강문봉 외 18인 공역 (1999). Robert E. Reys et al. 저. *초등 수학 학습지도의 이해*. 서울 : 양서원.
- 강완 · 백석운 (1998). *초등수학교육론*. 서울 : 동명사.
- 강지형 · 김수환 · 라병소 · 박성택 · 이의원 · 이정재 · 정은실 (1999). *7차 교육과정에 의한 초등수학교육*. 서울 : 동명사.
- 교육부 (1996a). *수학 3-1*. 충남 연기 : 국정교과서주식회사.
- 교육부 (1996b). *수학 4-2*. 충남 연기 : 국정교과서주식회사.
- 교육부 (1996c). *수학 5-2*. 충남 연기 : 국정교과서주식회사.
- 교육부 (1997a). *수학과 교육 과정*. 서울 : 교육부.
- 교육부 (1997b). *수학 6-1*. 충남 연기 : 국정교과서주식회사.
- 교육부 (2000). *수학 2-가*. 서울 : 대한교과서주식회사.
- 구광조 · 오병승 · 류희찬 공역 (1992). NCTM 저. *수학교육과정과 평가의 새로운 방향*. 서울 : 경문사.
- 김명숙 외 8명 (1999). *국가수준 교육성취도 평가 연구 II : 사회·수학 영역 예비 문항 개발 및 현장 적용 연구*. 서울 : 한국교육과정평가원.
- 김성숙 · 유준희 · 서동엽 · 이춘식 · 임찬빈 (1999). *제 3차 수학·과학 성취도 국제비교 반복 연구 (TIMSS-R) 국내 평가결과 분석 연구*. 서울 : 한국교육과정평가원.
- 김연식 · 김홍기 (1996). *중학교 수학 2*. 서울 : 두산동아.
- 김재춘 · 부재울 · 소경희 · 채선희 (2000). *교육과정과 교육평가*. 서울 : 교육과학사.
- 김정호 외 10명 (1999). *제 7차 교육과정에 따른 성취기준과 평가기준 개발연구—초등학교 1, 2학년—*. 서울 : 한국교육과정평가원.
- 박정 · 김성숙 · 홍미영 (2000). *제 3차 수학·과학 성취도 국제비교 반복 연구(TIMSS-R) 국내 평가결과 분석 연구 II*. 서울 : 한국교육과정평가원.
- 서동엽 (1999). *증명의 구성 요소 분석 및 학습 지도 방향 탐색—중학교 2학년을 중심으로—*. 서울대학교 교육학박사학위논문.

- 서동엽 (2000). 제 3차 수학·과학 성취도 국제비교 반복 연구에서 우리 나라 중학교 2학년 학생들의 수학 성취도 국제 비교. *우리 나라 중학생의 수학·과학 성취 결과, 국제 수준은 어떠한가?* 서울 : 한국교육과정평가원. pp.25-66.
- 우정호 (1994), 증명 지도의 재미미. *대한수학교육학회 논문집*. 제 4권. 제 1호. pp. 3-24.
- 우정호 (1998). *학교 수학의 교육적 기초*. 서울 : 서울대학교출판부.
- 이홍우 (1995). *증보 교육과정탐구*. 서울 : 박영사.
- 황혜정 (2000). 2000년도 국가수준의 중·고등학교 수학과 교육성취도 평가 연구. *대한수학교육학회지 수학교육학연구*. 제 10권 제 2호. pp.161-182.
- Hanna, G. (1993). The Validity of International Performance Comparisons. Mogens Niss(ed.). *Investigations into Assessment in Mathematics Education*. Dordrecht : Kluwer Academic Publishers. pp.245-252.
- Krutetskii, V. A. (1976). *The Psychology of Mathematical Abilities in School Children*. The University of Chicago Press.
- Mullis, Ina V.S. et al. (2000). *TIMSS 1999 : International Mathematics Report*. Boston : The International Study Center.
- Niss, Mogens (1993). *Assessment in Mathematics Education and Its Effects: An Introduction*. Niss, Mogens(ed.). *Investigations into Assessment In Mathematics Education: An ICMI Study*. pp.1-30.
- Robitaille, D. F. & J. S. Donn (1993). *TIMSS : The Third International Mathematics and Science Study*. Mogens Niss(ed.). *Investigations into Assessment in Mathematics Education*. Dordrecht : Kluwer Academic Publishers. pp.229-244.

<ABSTRACT>

Analysis on the Practices of Mathematical Assessment

Seo, Dong-Yeop¹⁵⁾

The aim of the study is to explore some problems that we have to solve to execute assessments effectively and in agreement with the objectives of them. We analysed the practices of some assessments including our national assessment of educational achievement and the third international mathematics and science study with focussing on the frames of assessments, the analyses of results, and the items presented in the assessments. The results of the study are the following. Firstly, we need to make the frame of assessment to agree with the objectives of assessment and to reflect the characteristics of the item related to a few areas. Secondly, we need to analyse the results of assessment with reflecting the frames of assessment. Thirdly, we need to discuss more concretely on the level and presentation of items including the order of conditions to need to solve the items. And lastly, we need to minimize the difference caused by the variations of translation in the international assessments.

15) Chuncheon National University of Education (339 Seoksa-Dong, Chuncheon-City, Kangwon-Do, 200-703, Korea.
Tel: 033-260-6452; E-mail: dseo@cnue.ac.kr)