Journal of the Korean Association for Science Education
Journal homepage: www.koreascience.org

과학 교육 평가에서 나타나는 고등학생들의 성취 불일치 사례 – 정의적 영역 검사 도구를 중심으로 –

정수임, 신동희*
이화여자대학교

Cases of Discrepancy in High School Students' Achievement in Science Education Assessment: Focusing on Testing Tool in Affective Area

Sue-Im Chung, Dong-Hee Shin*
Ewha Womans University

ARTICLE INFO

Article history:
Received 31 July 2017
Received in revised form
9 August 2017
4 September 2017
Accepted 7 September 2017
Keywords:
cognitive achievement, affective achievement, cases of discrepancy, response bias, self-reported testing tool, social desirability, science education assessment

ABSTRACT

This study analyzed some of the discrepancies in quantitative and qualitative data focusing on cognitive and affective achievement in science education. Academic and affective achievement score of 308 high school students were collected as quantitative data, and 33 students were interviewed for qualitative data. We examined the causes and types of discrepancies in terms of testing tools. As a result from quantitative data, there were a large number of students with a big difference between subjects in cognitive achievement, and constructs in affective achievement. More than 20% of the students did not match tendency between achievements in two areas. Through interviews, some examples such as intentional control of science learning for future study and careers, different responses by differences in perception between school science and science, appeared. A comparison of quantitative data by testing tool between qualitative ones and interviews showed conflicting result, where most students evaluated themselves differently from their own quantitative data. That is due to the students' interaction with the testing tools. Two types of discrepancy related to testing tool are found. One is 'the concept difference between the item developer and students,' the other is 'the difference between students' exposed response and their real mindset.' These are related to the ambiguity of the terms used in the tool and response bias due to various causes. Based on this study, an effort is required to elaborate the testing item that matches students' actual perception and to apply students' science learning experience to testing items.

Ⅰ. 서론

우리나라가 1995년 TIMSS(Trends in International Mathematics and Science Study)와 2000년 PISA(Programme for International Student Assessment)를 시작으로 국제 학업 성취도 평가에 참여한 이후 학생들의 성취 결과는 국가 단위 교육 정책의 효과와 현황을 가늠하는 척도가 되어오고 있다. 국제 학업 성취도 결과에서는 학업 성취 뿐 아니라 정의적 영역을 포함하는 교육 맥락 변인이 함께 분석되면서 과학 교육의 성과와 문제점을 다방면으로 파악하여 개선할 수 있는 기초 자료가 되었다. 우리나라에서 TIMSS와 PISA가 시행되었던 과정에서 끊임없이 제기되었던 높은 학업 성취와 낮은 정의적 성취[1] 간의 불일치(Kwak, 2017; Lee, 2016; Choe *et al.*, 2013; Cho *et al.*, 2012; Lee, Sohn, & No, 2007)는 최근에 발표된 TIMSS 2015와 PISA 2015의 분석 결과에서도 이어졌다(Ku *et al*., 2016a; Sang *et al*., 2016). 높은 학업 성취도에 비해 상당히 낮은 정의적 성취 현상은 대만이나 일본(Kim & Cho, 2013)과 같은 동아시아 국가에서도 공통으로 나타나, 아시아 특유의 문화적 정체성과 학습에 대한 문화적 맥락, 높은 수준의 학습 내용 등을 원인으로 제시한 연구도 발표되었다(Martin, Mullis, & Foy, 2012; Mullis, Martin, & Foy, 2012). Kim *et al*.(2009)은 각 국가를 하나의 단위로 설정해 산출한 PISA와 TIMSS의 성취 결과로부터 대체로 학업 성취가 높은 학생들이 낮은 정의적 성취를 보이는 불일치 현상을 보고했다. 이는 학업 성취가 높은 국가의 학생들이 그렇지 않은 국가에 비해 정의적 성취가 전반적으로 낮았다는 점에서 우리나라의 경우와 잘 들어맞는다. 그러나 우리나라 학생들끼리 비교한 국가 내 분석에서는 인지적 성취와 정의적 성취가 다른 국가들에 비해 높은 상관이 있었고(Kim *et al*., 2009; Ku *et al*., 2016a), 이로부터 낮은 정의적 영역을 긍정적으로 개선하여 두 영역의 성취를 균형 있게 향상시켜야 할 필요성이 도출되었다. 인지적 성취와 정의적 성취의 관계에 대한 연구는 정의적 특성이 인지적 성취에 영향을 주는 교육 맥락 변인이라는 입장에서 두 영역의 관계에 대한 연구(Kim, Kim, & Park, 2014; Park, 2007; Jűrges, Schneider, & Bűchel, 2005), 학습 경험과 정의적 영역 특성의 관계에 대한 연구(Lee & Kim, 2004; Osborne, Simon, & Collins, 2003), 학업 성취에 영향을 미치는 정의적 영역 내의 세부 구인인 흥미, 가치, 효능감 등을 다룬 연구(Kim & Seo, 2011; Seo, Choi, & Kim, 2007; Shen & Pedulla, 2000) 등으로 진행되었다. 이들은 학업 성취와 정의적 영역의 변인들 간에 밀접한 관계가 있으며 정의적 영역의 변인들 간에도 서로 높은 상관이 있음을 지시하고 있다. 또한 학령이 올라갈수록

* 교신저자 : 신동희 (donghee@ewha.ac.kr)
http://dx.doi.org/10.14697/jkase.2017.37.5.891

정의적 영역의 변인에 대하여 부정적 응답을 하는 것이 공통 현상으로 나타났다.

인지적 성취와 정의적 성취가 국가 간 분석과 국가 내 분석 결과에서 다른 경향을 보이는 이유를 주목할 만하다. 선행 연구 결과들은 인지적・정의적 성취의 정적 관계를 예측하지만 국가 간에는 학업 성취도가 높은 국가의 정의적 성취가 낮아서 오히려 부적 관계를 나타낸다(Kim *et al*., 2009; Ku *et al*., 2016a). 우리나라 학생들이 다른 나라에 비해 정의적 성취가 낮은 이유에 대해서 국내 연구자들은 높은 성취 수준에 비교되는 자신감 부족(Kim *et al*., 2012), 동양 문화권의 특색(Park *et al*., 2004), 중간 반응 선호 경향(Shin & Sohn, 2014) 등을 제시했다. 그러나 국가 간 두 영역의 성취 경향을 비교해서 나타난 불일치 현상을 해석하는 것은 쉽지 않다. 우선 우리나라와 일본, 대만 등 동양권 나라가 포함된 국가들은 높은 학업 성취도에 비해 정의적 성취가 낮다. 이들의 불일치 결과가 다른 국가들에 비해 뚜렷하다면 통계 값에 우세하게 반영되어 전반적 경향으로 나타날 수 있다. 이 경우 이들 나라의 사회 문화적 맥락을 포함한 특이성과 학생들의 반응 양식 간 관계를 파악해야 이 현상을 이해할 수 있다. 한편 학업 성취가 높은 국가에서 과학에 부정적으로 반응하는 현상에 대해서는 성취 수준과 맥락이 다른 학생들이 학습에서 겪는 경험과 그 속에서 느끼는 정서의 차이를 검토하여 대처할 필요가 있다. 두 요인은 인지적・정의적 불일치 현상에 어느 정도 영향을 주는 요소이고, 불일치 현상을 이해하고 균형 잡힌 성취의 방향을 모색하는 연구 문제로서 가치가 있다. 특히 전자의 경우 학생들이 속해 있는 사회 문화적 상황에 따른 영향을 최대한 통제할 수 있는 측정 도구를 마련하거나 반대로 다른 문화권의 결과를 함께 비교하고 해석할 수 있도록 보정하는 장치가 필요하다. 따라서 본 연구에서는 인지적・정의적 불일치 현상을 해석해서 그 원인을 찾아가는 첫 번째 단계로서 측정 도구와 상호작용하는 학생들의 반응을 살펴보았다.

일반적으로 정의적 영역 특성을 측정하기 위해서는 자기보고식 검사 도구를 사용한다. 자기보고식 평가는 참여자의 주관적 자기 보고에 의존하기 때문에 반응의 진실성 여부에 따라 검사 결과의 타당도와 신뢰도에 영향을 주고 반응의 경향성, 편향성, 정확성 측면에서 문제점이 제기되었다(Bae, Lee, & Ham, 2015; Song, 2010). 연구자들은 오래 전부터 응답자의 진실한 답변을 얻기가 쉽지 않음을 인지하고 있었고 이러한 문제들을 반응 왜곡(response bias)으로 정의했다(Cronbach, 1946; Schuluf *et al*., 2008). Messick(1991)은 이를 다시 검사에 상관없이 개인적 특성으로 나타나는 반응 양식(response style)과 검사 상황에 따라 발생하는 현상인 반응 세트(response set)로 분류했다. 반응 왜곡의 대표적 사례로 사회적으로 인정받으려는 방향으로 응답하려는 사회적 바람직성(social desirability)이 있는데 이는 응답자의 반응에 체계적 오류를 일으켜 잘못된 결과를 도출하는 주요 오염원이다(Bae, Lee, & Ham, 2015; Crowne & Marlowe, 1960; Cronbach, 1946). 따라서 자기보고식 검사 도구를 사용하는 심리학 영역에서는 일찍이 반응 왜곡을 통제하려는 체계적 노력을 기울이고 있는데 사회적 바람직성을 감지하는 척도를 개발하고 활용하는 연구(Bae, Lee, & Ham, 2015; Stöber, 2001; Paulhus, 1984; Crowne & Marlowe, 1960)와 측정 문항만을 통계적으로 처리해서 적합도 지수를 산출해서 활용하는 연구(Ferrando & Chico, 2001; Reise & Flannery, 1996) 등의 방향으로 수행되고 있다. 반응 왜곡을 감지하려는 시도는 특히 검사 결과의 파급력이 큰 분야에서 활발히 진행되고 있는데 심리 검사, 성격 검사나 상담 분야, 채용 및 선발을 위한 인적성 평가 등으로 점차 확대, 실용화되고 있다(Au, 2007; Son, Cha, & Kim, 2007; Stephen, 2000). 지금까지 과학 교육 학계에서는 정의적 영역 관련 자기보고식 평가 도구를 개발하거나 타당화하는 연구는 많았지만, 측정 대상을 잘못 묘사해서 결과를 오염시키는 체계적 오류로서의 반응 왜곡에는 크게 관심을 보이지 않았다(Shin & Sohn, 2014; Sohn, 2017).

학생들이 의식적이든 무의식적이든 솔직하게 반응하지 않는 이유는 앞서 사례로 든 사회적 바람직성 이외에도 여러 가지 가능한 이유들이 열려 있다. 한국의 사회 문화적 특이성이 집단의 반응에 영향을 줄 수도 있고, 특정 맥락에 관계없는 개인이 지니는 보편적 특성이 반응에 묻어나올 수 있으며, 검사의 상황에 따라 일관성이 없는 선택을 할 수도 있다. 어떠한 경우가 되었든 검사 도구 개발자가 왜곡 가능성이 있는 요소나 문화적 특이성을 사전에 인식한다면 반응 왜곡을 통제하는 도구를 제작하거나 반대로 왜곡을 감지함으로써 사회 문화적 특이성이나 개인적 특성, 반응 세트(response set) 상황을 해석할 수 있는 검사 도구를 연구 목적에 따라 개발할 수 있을 것이다. Abd-El-Khalick *et al.*(2015)은 아랍어를 사용하는 학생들의 과학적 태도를 측정하는 도구를 개발하면서 중동이나 아랍에서는 개인을 중시하는 서양과 달리 가족의 영향을 많이 받는 동양 문화권의 차이에 대한 보다 심층 연구가 필요하다는 의견을 제시했다. 서양에서 개발된 도구를 타 문화권에서 사용할 경우 단일 차원에서 비교할 수 없는 차이가 나타날 수 있다. 이를 반응 왜곡 차원으로 볼 것인지 문화적 차이에 의한 실재(實在)로 볼 것인지 단순히 판단하기 어렵다. 문화 비교 측면의 연구(Dudley *et al.*, 2005; Keillor, Owens, & Pettijohn, 2001), 한국 문화 특성에 대한 연구(Bae, Lee, & Ham, 2015; Kim, 2010), 다양한 맥락과 언어의 차이를 고려한 국제 비교 평가 연구(Chi, 2011; Murayama *et al.*, 2009; Suzuki & Ponterotto, 2007) 등을 종합적으로 분석할 필요가 있다. 아울러 이제까지 과학 교육에서 평가 도구 관련 문제점으로 지적되어 왔던 측정 개념의 명료성 부족과 이론적 기초가 명확치 않은 구인(Chung & Shin, 2016; Krynowsky, 1988; Koballa, 1988)도 주관적이거나 일관성이 부족한 반응을 초래하는 원인으로 꼽을 수 있다. 검사 도구가 지닌 체계적 오류는 응답자의 반응을 과장하거나 축소하는 기술적 문제 이외에도 측정하고자 하는 변수들 간의 관계에 영향을 미칠 수 있다는 점(Ganster, Hennessey, & Luthans, 1983)에서 인지적・정의적 성취의 불일치 사례를 포함하여 과학 교육 평가에서 나타나는 여러 가지 불일치 사례에 대한 원인으로 검토해 볼 가치가 있다.

한편, 인지적 성취에 비해 정의적 성취가 현저히 낮은 현상의 원인은 온전히 도구의 문제로만 돌릴 수 없는 명백한 현상이기도 하다. 상급 학교로 갈수록 과학에 대한 부정적 반응이 동서양의 공통 현상으로 나타나듯이(Koballa, 1988; Schunk & Pajares, 2009; Kwak, 2017) 학습 내용과 수준이 높아질수록 구체적, 경험적 수준에서 이해 가능한 영역이 감소하고 학생들은 실패와 좌절을 보다 많이 겪게 된다. 결국 불일치의 문제는 학생의 학습 경험 속에서 찾아야 한다. 학령에 따른 종단 연구에서 나타난 현상을 동일 연령의 학생들에게 적용

1) PISA와 TIMSS에 포함된 정의적 특성들의 성취를 의미하며 TIMSS 2003 결과보고서 이후 PISA와 TIMSS 평가에서 사용됨(Choe *et al.*, 2013).

하기 위해서는 학습 맥락과 성취 수준이 서로 다른 학생들이 경험하는 학습과 그 속에서 체감하는 정서와 느낌을 잘 살펴야 하며 이는 각자의 특성에 맞게 특화된 학습을 제공할 수 있도록 한다. Costa(1995)와 Aikenhead(2001)는 질적 면담을 통해 학교 과학에 적응해서 성취하는 양상에 따라 학생들을 몇 가지 유형으로 분류했다. 이들이 분류한 성취 유형별 특성은 학생이 처한 일상 문화와 과학 문화의 차이를 드러내 주었고 그 둘의 경계를 넘는 학습의 성공이라는 비전을 제시했다.

본 연구에서는 인지적 성취와 정의적 성취 수준에 따라 네 유형으로 분류한 학생들의 학업 성취도와 자기보고식 도구로 측정한 정의적 성취도, 과학 학습 경험에 대한 면담 자료를 분석했다. 그 결과 인지적·정의적 성취에서 불일치 사례가 나타남을 확인했고 그 유형과 원인을 검사 도구의 측면에서 살펴보았다. 본 연구의 연구 문제는 다음과 같다.

첫째, 양적으로 측정되는 인지적 성취와 정의적 성취를 중심으로 한 불일치 사례는 무엇이 있는가?

둘째, 학생이 검사 도구에 반응해서 생산한 양적 자료와 면담을 통해 이야기한 질적 자료에서 나타나는 불일치 사례는 무엇인가?

셋째, 검사 도구와 면담 결과에서 나타나는 불일치 사례의 유형과 원인은 무엇인가?

II. 연구 방법

1. 연구 대상

연구 대상은 경기도 중소 도시 소재 인문계 남녀 공학 고등학교 2학년 학생 308명으로, 이들에게 과학의 정의적 영역 특성 및 과학 학습 현황에 대한 설문 조사 실시 후 2016년 1학기 과학과에 해당하는 물리Ⅰ, 화학Ⅰ, 생명과학Ⅰ, 지구과학Ⅰ의 학기말 성취도 결과를 제공받았다(Table 1). 정의적 특성의 성취도와 학업 성취도를 분석하여 면담 대상자를 33명 선정하고, 학업 성취도의 제공을 포함하여 연구에 동의한 308명의 자료만 분석에 사용했다. 연구에 참여한 학생들이 속한 학급 형태는 인문 과정 2학급, 자연 과정 7학급, 과학 중점 과정 2학급이었다. 해당 연구 기간 동안 인문 과정 학생은 주 3시간 실시한 지구과학Ⅰ과목, 자연 과정 학생은 10시간, 과학 중점 과정 학생은 11시간에 해당하는 과학 교과 4과목의 학기말 성적을 T점수로 환산하여 인지적 영역 성취도로 제공했다. 정의적 영역 성취도는 과학이 주 영역이었던 PISA 2006에서 사용한 과학에 대한 태도 중 26문항을 선정하여 조사했다. 인지적·정의적 성취도를 수집한 후 정의적 성취도와 인지적 성취도를 기준으로 학생들을 4유형으로 분류했다. 이때 성취도 점수를 25%씩 4구간으로 나눈 분류 지표로서 분위를 정의해 사용했는데 상위 25%는 4분위, 최하위 25%는 1분위에 해당한다. 이들 중, 인지적 성취도와 정의적 성취도가 높고 낮음에 따라 분류된 A, B, C, D유형별 학생을 33명 선정하여 면담을 진행했다.

2. 연구 절차

국제 학업 성취도 평가인 PISA 2015에서는 우리나라 학생들의 정의적 영역 성취도가 PISA 2006에 비해 다소 개선되었지만, 인지적 성취도에 비해 정의적 성취도가 현저히 낮은 불일치 현상이 여전히 계속되고 있다(Ku *et al.*, 2016a; Choe *et al.*, 2013; Park, 2008). 학생들이 과학에 대해 지니는 긍정적인 정의적 특성은 이후의 학습을 지속시켜 평생 학습으로 나아갈 수 있는 동력이 된다(Lee, Sohn, & No, 2007)는 점에서 두 영역의 불일치 양상을 분석하여 그 원인과 대책을 찾아야 하는 필요성이 드러난다. 본 연구에서는 과학 교육 현장에서 나타나는 인지적 성취도와 정의적 성취도의 불일치 사례를 양적 자료와 질적 자료 분석을 통해, 교실 수준에서 나타나는 불일치 현상을 학생들의 과학 학습 경험의 맥락에서 조명했다. 이 연구는 2016년 9월부터 2017년 2월까지 수행되었으나, 분석 자료 중 하나인 인지적 성취도는 2016년 1학기말 성적을 사용했다. 이로 인해 학생들의 정의적 성취도와 인지적 성취도의 시점은 약 3개월 정도 차이가 난다. 학생들은 과학 교과를 이루는 물리Ⅰ, 화학Ⅰ, 생명과학Ⅰ, 지구과학Ⅰ의 4과목에 해당하는 1학기말 성적을 연구자에게 제공할 것과 정의적 영역 및 과학 학습 현황에 대한 설문에 참여한다는 동의서를 보호자의 동의하에 제출했다. 11학급 342명에게 배부했으나, 동의서를 제출하고 성실하게 응답한 학생 308명의 자료만 분석에 포함시켰다. 과학 교과 4과목의 학기말 성적은 1차 지필 평가와 2차 지필 평가를 비롯하여 수행 평가 점수가 합산된 점수이며, 각각 평균과 표준 편차가 달라 원점수로는 의미 있는 비교가 불가능해서 T점수로 환산한 후 각 과목의 평균 및 4과목 평균을 구했다. 정의적 영역 특성에 대한 설문지는 과학이 주 영역이었던 PISA 2006의 문항 중 과학에 대한 태도 문항의 일부였던 가치, 자아 개념, 과학에 대한 즐거움, 외적 동기에 대한 26문항으로 구성했다(Lee, Sohn, & No, 2007; Lee, Park, Sohn, & Nam, 2007). 설문지에는 정의적 영역의 문항 외에도 과학 학습 현황 및 반응 경향에 대한 문항 16개도 함께 조사되었다.

이로써 불일치 사례 탐색을 위한 양적 자료를 얻을 수 있었는데, 이들 자료를 SPSS 21.0에 의해 기술 분석 및 평균 분석, 분산 분석, 공분산 분석, 상관 분석 등의 통계로 처리했다. 수량화된 인지적 성취와 정의적 특성 점수의 평균을 오름차순으로 배열했을 때, 하위 25%까지는 1분위, 25% 초과 50% 이하는 2분위, 50% 초과 75% 이하는 3분위, 75% 초과인 상위 25%는 4분위로 분류하여 두 영역에서 나타

Table 1. Research subjects

분석 대상		내용
양적 자료	정의적 영역 성취도 검사 및 설문 조사	- 고등학생 2학년 308명(남 177명, 여 131명) - 정의적 영역 특성 검사지(정의적 성취도 26문항 포함 42문항)
	인지적 영역 성취도	- 308명의 과학 4과목(물리Ⅰ, 화학Ⅰ, 생명과학Ⅰ, 지구과학Ⅰ) 1학기말 성적 - 표준화 점수 활용(T점수 환산)
질적 자료	학생 면담	- 33명(남 20명, 여 13명) - 정의적·인지적 성취도에 따라 4유형 분류

나는 불일치의 정도를 분위의 차이로 나타냈다. 두 영역의 성취가 모두 낮거나 높은 학생은 각각 A형과 C형으로 분류하고, 인지적 성취는 높은데 정의적 성취가 낮은 B형, 인지적 성취는 낮고 정의적 성취가 높은 D형 학생으로 분류하면서 일치 유형과 불일치 유형이 나타났다. 네 유형에 속하는 학생 중 33명의 면담 대상자를 학부모의 동의를 얻어 선정했고 면담에 참여한 학생은 스스로 정한 가명을 사용했다. 면담을 담당한 연구자는 방과 후나 점심시간을 이용하여 33명을 1명 혹은 2명으로 나누어 총 22회 면담을 실시했으며 짧게는 27분, 길게는 82분으로 1인당 평균 면담 시간은 약 38분 정도 소요됐다. 학생들은 면담 전날 면담 내용에 대한 질문지와 연구 동의서를 사전에 받아 보호자의 동의서를 면담 당일 제출했고, 녹음된 면담 내용은 전사되어 NVivo 11.0 프로그램을 사용해 부호화되었다. 1차 주기의 부호화 과정에서 연구자는 시간 순으로 전사된 내용을 부호화했으므로 한 학생이 언급한 같은 의미의 진술은 반복 집계되었다. 후에 연구자들의 협의로 개별 부호들을 큰 단위로 묶어 나가는 과정을 통해 반복이나 중복 사항을 제거했다. 1차 부호화 과정은 가능하면 학생들이 쓴 생생한 말을 명사형으로 요약하거나 자료에 행동을 포함한 '~하기(함)' 등으로 표현했는데 "성적은 운에 맡기는 사다리 타기", "학교 과학은 수능 보는 도구", "벽보고 혼자 떠들며 공부하기" 등이 그 예다. Strauss(1987)와 Charrmaz(2006)는 참여자의 말을 직접 인용하여 부호화하는 방법은 의미를 구체화하고 요약하는 것이며 행위자들의 견해와 행동의 의미를 보존하는 의미가 있다고 했다. 행동으로 표현한 부호는 상황이나 문제에 대해 취한 행동, 상호작용, 감정을 조사하는 연구에 적합해서(Corbin & Strauss, 2008), 참여자들의 전술과 전략을 분석할 수 있다. 이런 현장 부호화와 과정 부호화 방법은 참여자의 관점을 적극 수용하고 반영했다는 점에서 의미가 있지만, Saldaña(2009)는 자료에 대한 연구자의 관점을 제한할 수 있으므로 더 개념적이고 이론적인 견해를 연구자의 관점에 포함해 부호화하도록 권장했다.

연구자 2인은 학생들의 목소리를 보존하여 부호화한 1,212개의 진술을 74개의 개념과 17개의 하위 범주로 묶고 다시 8개로 범주화하는 협의 과정을 거쳤다. 이때 각 부호들의 유사성과 차이점을 비교하여 면밀히 검토하면서(Corbin & Strauss, 2008) 쪼개지고 분열된 자료를 재조직했다. 속성이나 범위로 설정할 수 있는 진술을 하나의 개념으로 묶는 방법이 유용하게 활용되었다. 예를 들어 '어린 시절의 과학 경험'이라는 개념은 '긍정적', '부정적', '영향 없음', '긍정 후 부정' 등의 범위가 나타났으므로 이를 하나의 개념으로 설정했다. 이렇게 범주화한 학생들의 과학 학습 경험을 인지적·정의적 성취가 일치하거나 일치하지 않는 네 유형 학생별로 다시 분석했다. 그 결과는 각 유형 학생들의 특성으로 요약되었는데, 과학 학습 평가 상황에서 영역별 성취가 일치하거나 불일치한 학생들의 특성이 그들의 학습 경험으로부터 도출되었다.

3. 분석 자료

본 연구에서는 과학 교육 평가 현장에서 인지적 성취와 정의적 성취가 일치하지 않는 경우를 다양한 측면에서 조명해 보기 위해 양적 자료와 질적 자료를 분석했다. 양적 자료로는 학업 성취도 점수, 정의적 특성 및 학습 경험 설문지를 각각 인지적 성취와 정의적 성취의 분석 자료로 사용했고, 질적 자료로는 학생들과의 면담 내용을 전사하여 활용했다. 따라서 연구자들은 정의적 특성 및 학습 경험 설문지와 면담용 반 구조화된 질문지 등 2종을 개발했다(Table 2, Table 3). 먼저 양적 자료를 얻기 위한 과학 관련 정의적 특성 및 학습 경험 설문지(Table 2)는 과학이 주 영역이었던 PISA 2006의 과학에 대한 태도 문항에서 26문항을 선정했다. 연구자들은 문항 선정을 위해 PISA 2006 문항과 함께 TIMSS 2011의 정의적 문항을 비교, 검토했다. TIMSS 2011은 과학에 대한 자신감과 함께 흥미를 나타내는 내적 동기와 가치를 인식하는 외적 동기 등 3개의 구인을 설정했고(Kim *et al.*, 2012), 이와 유사하게 PISA 2006은 과학에 대한 일반적 가치와 개인적 가치, 자아 개념, 즐거움, 외적 동기에 대한 문항을 포함했다(Lee, Sohn, & No, 2007). TIMSS 2011의 자신감 척도와 내적 동기 척도는 각각 PISA 2006의 자아 개념과 즐거움에 해당하지만, PISA 2006이 일반적 가치와 개인적 가치, 외적 동기(도구적 가치)로 가치 척도를 세분화한 반면, TIMSS 2011은 이를 외적 동기로 간략화 했다. 이에 본 연구는 구인을 보다 세분화한 PISA 2006 문항을 선택했다. 또한 한국 사회는 직업이나 진로 선택에 사회적 압력이 강한 편으로 개인의 흥미와 선택을 강조하는 서양의 경우와 차이가 있을 것(Abd-El-Khalick *et al.*, 2015; Cho, 2003; Heine *et al.*, 2002; Myeong & Crawley, 1993)으로 생각하여 배경 변인으로서 주관적 규범에 대한 4문항(Myeong & Crawley, 1993)을 포함했다. 그밖에 학생들이 중요하게 생각하거나 선호하는 과목에 대한 정보에 대한 6문항, 문항에 대해 평소의 생각보다 높이거나 낮추어 왜곡하려는 경향이 있는지에 대한 6문항을 함께 조사했다. 자기보고식 검사에 내재한 취약점으로 참여자의 반응 세트에 따라 부정이나 긍정 왜곡이 일어날 수 있음을 감안하여 과학에 대한 정의적 영역에도 왜곡 경향이 나타나는지 분석했다.

Table 3은 학생들이 과학과 과학 학습 경험에 대해 어떻게 생각하고 있으며 과학 학습의 어려움에 어떻게 대처하는지, 설문 문항에

Table 2. Contents of questionnaire

항목	내용	형태	해당 문항 번호
과학에 대한 정의적 특성	과학 과목에 대한 자아 개념	리커트 4점 척도	1~6
	내적 동기(즐거움)		7~11
	가치(일반적, 개인적, 도구적-외적 동기)		12~26
주관적 규범성	가족, 친구들이 과학 관련 직업을 갖도록 도움을 주는지 여부 가족, 친구들이 원하는 직업을 갖고 싶은지에 대한 자신의 생각	리커트 4점 척도	27~30
과학 학습 상황	학교 과학 이외에 받는 교육 선호 과목 및 중요하다고 생각하는 과목	서술식	31~36
반응 경향	자아 개념, 선호도, 가치에 대한 반응의 왜곡 여부	선택형	37~42

반응하면서 무엇을 느끼고 혼란스러웠는지를 면담하기 위한 질문 내용이다. 이 질문들을 통해서 양적 자료에서 파악하기 어려운 질적 자료를 얻었다. 이 연구에 사용된 질문은 크게 4항목으로 구분되며 과학에 대한 인식, 성취에 대한 자기 진단 및 인식, 과학의 성취에 영향을 미치는 요인, 성취에 대한 전망 등이다. 특히 '과학에 대한 인식' 항목에서는 일상에서 생각하는 과학과 학교 과학을 분리해서 인식하고 있는지 먼저 확인했다. 이는 검사 상황에서 학생들이 자주 하는 질문에서 연구자들이 감지한 사항을 반영한 것이다.

2016년 9월부터 참여 학생들을 대상으로 Table 2의 설문 조사를 하던 중, 11학급 중 5개 학급에서 특정 문항에 대해 한 두건의 질문이 들어왔다. 첫 번째는 "과학은 물리, 화학, 생물, 지구과학 중 무엇으로 생각해야 하나"였고, 두 번째는 "과학은 학교에서 배우는 과목을 말하는 것인지 아니면, '그냥 과학'을 말하는 것인지 알려 달라"는 것이었다. Costa(1995)가 면담한 학생들은 학교 과학을 일상과 '다른 세계(another world)'로 인식했고, 이들이 경험하는 일상과 학교 과학의 괴리를 극복하기 위해서는 학생들이 과학 문화 속으로 옮겨가는 문화적 경계 넘기(crossing cultural borders)를 경험해야 한다(Aikenhead, 2001; Aikenhead & Jegede, 1999). 검사 문항에서도 과학과 학교 과학이라는 경계를 고려해야 함이 드러났고, 학생들이 느끼는 경계가 사회적 구성주의로 설명할 수 있는 것인지, 정의적 영역의 성취 결과가 이들 개념 간 불일치의 영향을 받을 가능성이 있는 것인지 점검해야 할 필요가 있었다. 과학 학습 경험에 대한 이야기를 나누는 동안 연구자는 면담 참여자에게 주관적 규범성(Table 2의 27번~30번)과 반응 경향(Table 2의 37번~42번)에 응답한 참여자 본인의 설문 결과를 보여주며 그렇게 응답한 이유에 대해서 떠올려 보게 했다. 자기보고식 검사에서는 실제 생각이나 솔직한 감정이 주변 사람들의 압력에 의해 다르게 표현되거나, 의식적 혹은 무의식적으로 자신을 포장하려는 반응 편파가 일어날 수 있다(Lee *et al.*, 2016; Bae, Lee, & Ham, 2015). 학생의 솔직한 마음과 밖으로 표현된 반응의 불일치는 검사에 내재한 한계일 수 있지만, 이 현상이 집단적 특성으로 나타날 경우 혹은 한 개인에게 특별히 가중되어 나타날 경우에는 검사 결과를 신뢰할 수 없으며 해석에 주의를 요한다. 따라서 검사 결과에 영향을 미칠 수 있는 가능한 요인을 점검함으로써 검사 도구를 정교화할 수 있는 시사점을 도출할 수 있다.

Ⅲ. 연구 결과 및 논의

본 연구는 인지적·정의적 성취를 중심으로 과학 교육 평가 현장에서 나타난 여러 가지 불일치 사례를 통해 불일치의 원인과 유형을 검사 도구의 측면에서 분석했다. 첫 번째 불일치 사례는 학생들의 학업 성취도 검사와 정의적 특성 검사에서 조사한 인지적 성취와 정의적 성취의 경향이 일치하지 않는 사례, 두 번째는 검사 도구로 측정한 결과와 면담 결과가 서로 일치하지 않는 사례다. Figure 1은 연구 문제들 간의 관계를 간략히 정리한 것으로, 각각의 불일치 사례가 첫 번째와 두 번째 연구 문제에 해당되고 불일치 사례가 나타난 원인을 검사 도구 측면에서 살펴본 세 번째 연구 문제는 검사 도구와 상호작용하는 학생의 반응을 두 가지 유형으로 제시했다.

1. 검사 자료 내 불일치 현상

가. 인지적 성취도 검사

일반 고등학교 학생들은 학년이 올라가면서 자신의 진로에 맞는 교육과정을 선택하게 된다. 그 과정에서 겪게 되는 큰 변화 중 하나는 선택하는 교육과정에 따라 사회나 과학을 이수하는 단위 수가 달라진다는 점이다. 연구에 참여한 학생들의 경우 1학년 때는 보통 교과의

Table 3. Questions for semi-structured interview

항목	질문
과학에 대한 인식	1. 내가 생각하는 '과학'은 어떤 것이라고 생각하나요? 2. 학교에서 배우는 과학 교과(물리, 화학, 생명과학, 지구과학)를 내가 생각한 '과학'과 비교해 보았을 때 일치하는 점은 무엇이고, 다른 점은 무엇인가요? 3. 내가 생각한 과학과 학교에서 배우는 과학 교과의 성질이 일치하거나 차이가 나는 점 때문에 학습에서 특별히 어렵거나, 쉽거나, 흥미가 생긴다거나 하는 느낌이 있다면 자유롭게 이야기해 주세요.
과학 관련 성취에 대한 자기 진단 및 인식	4. 제시되는 그래프에서 자신의 위치는 어디쯤이라고 생각하나요? 5. 나는 과학을 잘한다고 생각하나요? 그렇게 생각하는 이유는 무엇인가요? 6. 나의 과학 성적이 (높은/낮은) 이유는 무엇 때문이라고 생각하나요? 7. 나의 과학에 대한 정의적 성취(느낌이나 감정, 태도)가 (높은/낮은) 이유는 무엇 때문이라고 생각하나요? 8. 나의 경우, 과학 성적과 과학의 정의적 성취(느낌, 감정, 태도)는 어떤 관계가 있다고 생각하나요? 9. 앞으로 나의 과학 성적은 어떻게 될 것이라 예상하나요? 또한 과학에 대한 나의 정의적 성취(느낌, 감정, 태도)는 앞으로 어떻게 될까요?
과학에 대한 인지적, 정의적 성취에 영향을 미치는 요인 및 상황	10. 나의 과학 성적에 영향을 미치는 요인이나 상황이 있다면 자유롭게 이야기해 주세요. 11. 가족, 선생님, 친구 주위 사람들이 과거나 현재 지원하고 있는 도움이 특별히 과학 학습과 과학에 대한 느낌, 정서, 태도 등에 어떠한 영향을 미치는지 자유롭게 이야기해 주세요. 12. 우리나라의 사회적, 문화적 분위기가 특별히 과학 학습과 과학에 대한 느낌, 정서, 태도 등에 어떠한 영향을 미치는지 자유롭게 이야기해 주세요. 13. 참여자는 가족과 주위 사람들이 나에게 원하는 직업이나 일을 갖고 싶은지를 묻는 질문에 ( )라고 응답했습니다. 그 이유는 무엇인가요? 14. 가족과 주위 사람들의 의견이 나의 과학 성적이나 과학에 대한 느낌, 정서, 태도에 영향을 미칠 수 있다고 생각하나요? 15. 참여자는 반응의 왜곡 여부에 대한 질문에 ( )라고 응답했습니다. 그 이유는 무엇인가요?
인지적-정의적 영역 성취 지도(map)에서 자기 위치 이동에 대한 인식	16. 4번에서 제시된 그래프에서 앞으로 자신의 위치는 어떻게 될 것이라고 생각되나요? 그 위치를 향상(유지)하기 위해서는 어떤 노력을 해야 할까요? 17. 4번의 그래프에서 원하는 방향의 성취를 위해서 필요한 도움은 무엇이 있고, 이를 방해하는 요소는 무엇이 있을까요?

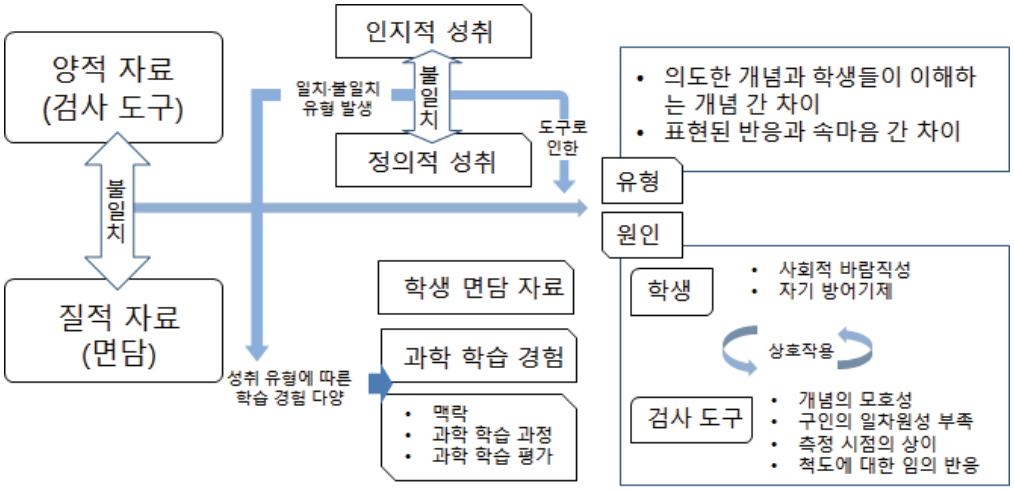

Figure 1. Summary of this study

일반 과목인 '과학'을 8단위 이수했다. 2학년이 되어 인문 사회 과정을 선택한 학생들은 생명과학 I 과 지구과학 I 을 각 3단위씩 총 6단위, 자연과정을 선택한 학생들은 물리 I , 화학 I , 생명과학 I , 지구과학 I 을 각 5단위씩 총 20단위 이수하게 된다. 자연 과정 학생들에게는 주당 4시간 수업했던 과학이 10시간으로 늘어난 셈이다.

> 상황이 많이 다르죠. 그때는 다 합쳐졌으니까 공부할 양도 되게 적어서 할 만 했는데, 이제는 네 과목으로 분리가 되고 양도 기하급수적으로 엄청 늘어나니까 이게 네 과목에 전부 신경을 쓰려 하니까 좀 머릿속에서 엉키는 것도 있고. 여기 조금하다가 저기 조금하다가 하나를 완벽하게 하다보면 또 여기를 할 시간이 별로 없고.
>
> (D유형 학생 김인정 면담 중)

학생들은 과학 수업 시수가 증가하면서 많아진 학습량뿐 아니라 동시에 4과목을 공부해야하는 어려움을 이야기했다. 분리된 4과목의 과학은 4과목의 평가를 동시에 받아야 함을 의미한다. Schunk & Pajares(2009)는 상급 학교로 갈수록 학습량이 늘어나고 추상적 내용을 설명식으로 학습하면서 실패와 좌절의 경험을 겪는 학생들의 자아효능감이 낮아지는 과정을 언급했다. Table 4는 학생들이 동시에 시험을 치룬 4과목 성적- 2016년 1학기 학기말 학업 성취도에서 산출한 과목 간 상관 계수를 정리한 것으로 모든 경우 유의 확률 $p<.01$ 수준에서 통계적으로 유의했다. 인문 사회 과정 학생들은 4과목 중 1과목인 지구과학 I 만 이수했으므로 분석에서 제외했다. 생명과학 I 이 화학 I , 물리 I , 지구과학 I 과 이루는 상관 계수가 각각 .878, .854, .832로 다른 과목 간 상관에 비해 고르게 나타났다. 지구과학 I 은 물리 I 과 .798, 화학 I 과 .797로 근소한 차이로 낮았지만, 전반적으로 4과목 간 학생들의 성취도는 높은 상관을 나타냈다. 대체로 남학생이 여학생보다 과목 간 상관이 작은 차이로 높았고, 학급 유형별 차이가 나타났으나 일관성 있는 경향을 보이지 않았다.

학생에 따라 성취도의 경향이 일치하지 않아 과목 간 성적 차이가 나타나는 경우를 Table 5에 정리했다. 앞서 정의한 상대적 서열을 기준으로 한 분위를 비교했을 때, 과목 간 분위 차가 크게 나타나는 학생들은 같은 과학 교과 내에서 과목에 따라 성취도의 불일치가 나타나는 학생들이다. 과목 간 분위 차가 가장 크게 나타난 과목은 물리 I 과 지구과학 I 으로 42.7%의 학생들의 과목 순위가 서로 다른 분포를 보였고, 가장 분위 차가 작은 과목은 화학 I 과 생명과학 I 이 34.3%였다. 과목 간 분위차가 1인 것은 최대 25% 이하의 순위 차이가 나타나는 경우로, 실제로는 분류 기준 근처에서 나타날 수 있는 근소한 차이가 포함된 값도 있으므로 해석에 유의할 필요가 있다. 그러나 분위 차가 2이상인 것은 과목 간 순위가 최소 25% 초과 50% 이하이거나 50% 초과의 차이를 보이는 경우이므로 불일치 정도가 상당히 크다. 6.8%의 학생들은 화학 I 과 지구과학 I 에서 25% 이상의 순위 차이를 나타냈는데, 이는 한 과목 점수가 다른 과목보다 현저히 높거나 낮게 나타남을 의미한다. 화학 I 의 석차가 지구과학 I 보다 높아 2분위 차를 나타내는 한 학생의 이야기다.

> 제가 원하는 걸 하기 위해서, 발판이 되어야 하는 걸 1순위로 잡으라

Table 4. Correlation coefficient between cognitive achievement of science subjects

과목	T점수로 환산한 학기말 학업 성취도의 과목 간 상관 계수*				
	성별		학급 유형별		전체 (N=251)
	남(N=152)	여(N=99)	과학 중점 과정(N=60)	자연 과정(N=191)	
물리I-화학I	.879	.873	.887	.868	.871
물리I-생물I	.874	.856	.845	.860	.854
물리I-지구과학I	.815	.832	.718	.817	.798
화학I-생명과학I	.881	.855	.903	.868	.878
화학I-지구과학I	.796	.785	.791	.802	.797
생명과학I-지구과학I	.829	.820	.778	.850	.832

*상관관계 전체가 모두 $p<.01$

Table 5. Number of students with a big difference in cognitive achievement between subjects

과목 간 분위* 차	과학 과목 간 학업 성취도 점수에서 상대적 차이가 나타나는 학생 수(%) (N=251)					
	물리I-화학I	물리I-생명과학I	물리I-지구과학I	화학I-생명과학I	화학I-지구과학I	생명과학I-지구과학I
1	92(36.6)	86(34.3)	96(38.3)	84(33.5)	83(33.1)	81(32.3)
2	5(2.0)	6(2.4)	11(4.4)	2(0.8)	16(6.4)	10(4.0)
3	0(0.0)	0(0.0)	0(0.0)	0(0.0)	1(0.4)	1(0.4)
계	97(38.6)	92(36.7)	107(42.7)	86(34.3)	100(39.9)	92(36.7)

* 분위는 학업 성취도 점수를 25%씩 4구간으로 나눈 분류 지표로 4분위는 최상위 25%, 3분위는 상위 75~50%, 2분위는 상위 50~25%, 1분위는 하위 25%를 나타냄.

> 하니까 그걸 1순위로 하다 보니 다른 건 떨어질 수밖에 없죠. 상대적으로 다른 걸 할 시간이 없는 거죠. 제 걸 하기에도 바쁜 거예요.
>
> (B유형 학생 김승주 면담 중)

> 생명과학 쪽에 흥미가 있거든요. 몸에 대해서 신비롭고 그런 거는 재밌게 공부를 할 수 있는데 물리랑 지구과학 같은 거는 전혀 흥미가 없어가지고 그게 잘 이해가 안돼요. 시험을 보려고 주입식으로 공부를 하는 거잖아요. 그래서 그걸 참고 공부를 해야 한다는 게 힘들어요.
>
> (C유형 학생 이가현 면담 중)

과학 과목 간 높은 상관(.797~.878)에서 볼 수 있듯이 성취도가 높은 학생들은 대개 과학 교과에 대한 성취도 경향이 비슷하여 한 과목에서 높은 점수를 받으면 다른 과학 과목에서도 높은 점수를 받는 편이다. 그러나 김승주 학생처럼 특정 과목에 대한 선택적 학습이 이루어지면 상관이나 경향을 따지는 것이 무의미해지며, 이가현 학생의 경우와 같이 과목에 대한 선호나 이해도에 차이가 생기면 인지적 성취도에서 과목 간 불일치 원인이 될 수 있다. 과학 교과가 4과목으로 분리되면서 두드러지게 된 학문 특성은 학생들에게 과목에 대한 선호를 불러일으킨다. 이는 학습을 촉진하기도 하지만 동기를 감소시키기는 원인이 되기도 한다.

나. 정의적 특성 검사

PISA 2006의 과학에 대한 태도 문항 중 자아 개념, 내적 동기(즐거움), 가치 등 세 가지 하위 구인을 학업 성취도 점수 분위에 따라 나타냈다(Table 6). 인지적 성취에 해당하는 학업 성취도가 높을수록 세 구인 모두 집단별 평균 차이가 유의 확률 $p<.01$ 수준에서 통계적으로 유의하며 2.4~3.1의 범위로 다양했다. 과학 과목에 대한 자아 개념은 학문적 능력에 대한 신념을 질문하고 있는데 가치(3.0)나 내적 동기(2.8)의 평균값보다 낮은 2.6으로, 리커트 1~4점 척도에서는 중간 정도에 해당한다. 자신의 능력을 비교적 낮게 평가하려는 경향은 성공적 학습을 방해하는 요소가 될 수 있다는 점에서 인지적 성취가 낮은 학생들의 결과를 주목할 필요가 있다. 특히 학업 성취도에 의한 정의적 성취도의 설명력인 상관비 $\eta 2$은 0.242이므로 둘 사이의 상관계수는 약 .49로 비교적 높은 상관이 있다. 한편 PISA 2015에서 우리나라는 정의적 영역의 지수인 과학의 즐거움, 자아 효능감, 도구적 동기 등에 의한 과학 성취도의 설명량이 차례로 14.6%, 7.1%, 4.7%로 나타났다. OECD 국가 평균은 이보다 낮아서 각각 9.4%, 6.0%, 1.5%로 나타났는데(Ku *et al*., 2016a), 이로써 우리나라가 다른 국가에 비해 인지적 성취도와 정의적 성취도의 상관이 보다 긴밀하게 관련됨을 알 수 있다. 본 연구에서는 독립 변인으로 인지적 성취도를 사용했다. Ku *et al*.(2016a)은 정의적 성취도에 의한 인지적 성취도의 설명량을 분석한 점이 다르지만, 그 결과는 두 영역이 어느 정도 상관이 있음을 시사하고 있다.

정의적 특성을 이루는 하위 구인 세 가지를 하나로 묶어서 각 영역의 합산 점수를 정의적 성취도로 나타낼 수 있을지 세 구인의 상관계수를 Table 7에 정리했다. 자아 개념, 내적 동기, 가치에 대한 Cronbach's α 계수는 각각 .914, .891, .941로 전체 문항에 대한 계수는 .956이다. 각 구인 간 상관 계수를 살펴보면, 자아 개념과 가치가 .559로 비교적 높은 상관관계, 자아 개념과 내적 동기(.710) 및 내적 동기와 가치(.708)는 높은 상관관계를 나타냈다. 상대적으로 상관 계수가 낮게 나타난 구인들은 자아 개념과 가치였는데, 학생들은 과학과 과학 과목을 얼마나 중요하고 가치 있게 생각하는지 여부와 자신의 학문적 능력에 대한 신념을 별개로 생각하려는 경향을 드러낸다. 대체로 남학생이 여학생보다 구인 간 상관이 높았고, 특히 여학생이나 인문 과정 학생의 경우 자신의 학문적 능력에 대한 확신과 과학을 가치 있게 여기는 마음은 각각 .446과 .397로 상대적으로 상관이 적었고 자연 과정 학생들의 경우(.507)도 마찬가지였다. Cho *et al*.(2012)은 진학이나 직업을 얻는 등 특정 목적을 위해 성취하려는 도구적 가치 인식의 증가는 오히려 내적 동기인 흥미나 즐거움이 낮아지는 원인으로 작용한다는 의견을 제시했다. 본 연구에서는 자아 개념과

Table 6. Science-related affective achievement of groups by the level of cognitive achievement

정의적 특성의 하위 구인	학업 성취도 점수 분위별 과학 관련 정의적 특성 점수 평균 (N=308)							
	1분위 (하위 25%)	2분위 (25-50%)	3분위 (50-75%)	4분위 (상위 25%)	평균	표준편차	F	$\eta 2$
자아 개념	2.4	2.5	2.7	3.0	2.6	0.61	17.62*	0.148
내적 동기(즐거움)	2.3	2.8	3.0	3.1	2.8	0.66	24.50*	0.195
가치	2.5	3.1	3.0	3.2	3.0	0.56	30.83*	0.233
평균	2.4	2.9	2.9	3.1	2.9	0.53	32.37*	0.242

*$p<.01$

Table 7. Correlation coefficient between sub-constructs in science-related affective characteristics

	과학 관련 정의적 특성의 하위 구인 간 상관 계수*					
	성별		학급 유형별			전체 (N=308)
	남(N=177)	여(N=131)	인문 과정(N=57)	과학 중점 과정(N=60)	자연 과정(N=191)	
자아 개념-내적 동기	.759	.613	.575	.671	.724	.710
자아 개념-가치	.605	.446	.397	.688	.507	.559
내적 동기-가치	.754	.613	.634	.617	.684	.708

* 상관관계 전체가 모두 p<.01
Cronbach's α(자아 개념)= .914, Cronbach's α(내적 동기)= .891, Cronbach's α(가치)= .941, Cronbach's α(전체 문항)= .956

가치의 상관이 보다 낮게 나타났다. Table 8에서 가치와 함께 묶여 있는 자아 개념(.507)과 내적 동기(.684)는 자아 개념과 내적 동기 간 상관관계(.724)보다 낮게 나타났다.

다. 인지적 성취도와 정의적 성취도와의 관계

Kim *et al.*(2009)은 TIMSS 2007과 PISA 2006의 과학 성취도와 정의적 영역 결과를 국가 내와 국가 간 상관 계수로 나타낸 바 있다. TIMSS 2007에서 인지적 성취도에 대한 자신감, 즐거움, 가치 지수의 상관 계수는 각각 .478, .391, .329로 나타났고, PISA 2006에서는 자아 개념, 즐거움, 도구적 동기와 인지적 성취도의 상관 계수가 각각 .375, .425, .254로 분석되었다. PISA 2006은 만 15세, TIMSS 2007은 중학교 2학년에 해당하는 학생들을 분석한 결과로 본 연구 대상자인 2016년 고등학교 2학년 학생들과는 시간적, 연구 대상의 특성 면에서 차이가 있다. 또한 국제 수준에서 합의된 교육과정 내용을 다루는 TIMSS와 실생활에 필요한 과학적 소양을 측정하는 PISA 성취 결과(Kim *et al.*, 2009)를 우리나라 고등학교에서 한 학기 동안 실시한 학업 성취도 결과와 직접 비교하는 데도 어려움이 있다.

Table 8은 본 연구에서 분석한 고등학교 2학년 학생들의 학업 성취도와 과학 관련 정의적 특성 검사 점수의 상관 계수를 정리한 것이다. 본 연구에서 2016년의 분석에 사용한 인지적 성취도는 학교 교육과정에서 이루어진 8번의 지필 평가와 각 과목의 수행 평가가 반영된 결과다. 국제 학업 성취도와 같이 특정 시점의 인지 수준을 1회 측정한 값이라기보다 평가를 대비한 일정 기간의 의도된 노력이 반영된 것도 다른 점이다. 인지적・정의적 성취의 상관은 자아 개념(.367), 내적 동기(.335), 가치(.255) 순으로 전체적으로 낮은 상관관계를 보였는데, 2006년과 2007년의 국제 학업 성취도 평가와 비교해서 시간, 대상, 검사 특성 등의 차이에도 불구하고, 어느 정도 일관성 있는 결과다. 특히 TIMSS의 가치 지수와 PISA의 도구적 동기에 해당하는 본 연구의 가치 영역에서 인지적 성취에 대한 상관은 자아 개념(자신감)이나 내적 동기(즐거움)에 비해 낮게 나타났다. 이는 두 국제 학업 성취도 결과와 같다. Table 6에서 학생들은 가치 점수(3.0)를 자아 개념(2.6)과 내적 동기(2.8)에 비해 높은 점수를 주었으나 Table 8에서 보듯이 인지적 성취와의 상관은 비교적 낮았다. 특별히 이 연구 대상의 대다수가 이공계 학생들이라는 점을 고려한다면, 학생들은 경험을 통해서 과학이 실제 내 삶에 중요하다고 인식하지만, 그러한 인식이 인지적 성취에 연결되는 정도는 학문적 능력에 대한 확신이나 즐거움보다는 덜하다는 점을 알 수 있다. Kwak(2017)은 TIMSS 2011에 비해 2015년에는 우리나라 학생들의 과학에 대한 도구적 가치 인식이 증가하고 있다는 자료를 제시했고, 국제적으로 과학에 대한 자신감과 가치 인

Table 8. Correlation coefficient between cognitive achievement and science-related affective achievement

정의적 특성의 하위 구인	학업 성취도 점수와 과학 관련 정의적 특성 검사 점수의 상관 계수					
	성별		학급 유형별			전체 (N=308)
	남(N=177)	여(N=131)	인문 과정(N=57)	과학 중점 과정(N=60)	자연 과정(N=191)	
자아 개념	0.492**	0.268**	0.101	0.464**	0.384**	0.367**
내적 동기(즐거움)	0.462**	0.205*	0.159	0.320*	0.363**	0.335**
가치	0.327**	0.236**	0.050	0.392**	0.239**	0.255**
평균	0.437**	0.278**	0.107	0.444**	0.344**	0.337**

*p<.05, **p<.01

Table 9. Distribution of students by the level of cognitive and affective area

		정의적 영역 점수의 각 분위별 해당 학생 수(%) (N=308)				계
		1분위 (하위 25%)	2분위 (25-50%)	3분위 (50-75%)	4분위 (상위 25%)	
인지적 영역	1분위(하위 25%)	48(15.6)	15(4.9)	7(2.3)	7(2.3)	77(25.0)
	2분위(25-50%)	14(4.5)	22(7.1)	24(7.8)	17(5.5)	77(25.0)
	3분위(50-75%)	11(3.6)	22(7.1)	24(7.8)	20(6.5)	77(25.0)
	4분위(상위 25%)	4(1.3)	16(5.2)	23(7.5)	34(11.0)	77(25.0)
계		77(25.0)	75(24.4)	78(25.3)	78(25.3)	308(100.0)

식이 높을수록 높은 인지적 성취도를 보인다고 보고했다. 한편 국가 간 분석을 실시한 결과, 이와 다르게 인지적 성취도가 높은 국가들의 정의적 성취도가 대체로 낮아서 부정적 상관관계를 나타냈다(Kim *et al.*, 2009). 인지적 성취도와 정의적 특성의 하위 구인들은 성별과 학급 유형별 집단 차이가 나타났다. 남학생(.437)이 여학생(.278)에 비해 두 영역의 상관이 상대적으로 높았고, 학급 유형에 따라서는 과학 중점 과정(.444)이 가장 상관이 높고, 자연 과정(.344), 인문 과정(.107)의 순으로 이어졌다.

학생별로 인지적・정의적 영역의 상관관계를 파악하기 위해 성취도를 오름차순으로 배열한 후 하위 25%부터 분위를 매겨 각 분위별 해당 학생 수를 정리하면(Table 9), 두 영역에 대한 분위 차가 3으로 가장 큰 경우는 11명(3.6%), 2인 경우는 51명(16.6%)에 해당해서 20% 이상의 학생들이 인지적 성취와 정의적 특성 간 불일치 경향을 보였다.

다음은 인지적・정의적 성취가 불일치하는 유형으로 분류한 학생들의 생각이다.

> 지금 흥미가 약간 떨어지는 과학 분야에 대해서도 점수가 낮게 나오는 편은 아니거든요. 싫어도 시험은 봐야 하니까 열심히 공부만 하면, 그리고 과학은 지식이잖아요. 지식이 성적과 관련이 있는데 그런 지식 면은 흥미와는 별도니까.
>
> (B유형 학생 안정현 면담 중)

> 과학에 대한 느낌이 아무리 좋아도 시험으로 보는 과학은 또 다르니까, 그게 자기 생각만의 과학이 아니라 딱 정해진 과학이기 때문에 별로 관계가 없다고 봐요.
>
> (D유형 학생 김인정 면담 중)

> 저는 과학에 대한 흥미는 많은데 그 성적 같은 것은 제가 다른 걸 공부하고 있어서 지금은 좀 멀리 하는 것 같아요.
>
> (D유형 학생 김준홍 면담 중)

안정현 학생은 과학의 지식과 흥미를 다른 차원으로 인식한다. 학업 성적이 우수한 편으로 이 학생의 공부는 흥미에서 시작되기보다 좋은 시험 성적을 얻으려는 동기에서 추진력을 얻는다. 김인정 학생은 학업 성적에 비해 정의적 성취가 높은 학생으로 '자기 생각만의 과학'과 '딱 정해진 과학'이라는 이분법을 과학에 적용했다. 자신이 일상 경험에서 흥미를 강하게 느끼는 과학과 학교에서 평가하는 정해진 과학은 다르기 때문에 과학에 대한 느낌이나 흥미는 성적과 관련이 없다고 생각했다. 결국 학교 과학은 흥미가 없어서 성적이 좋지 않지만, 이 학생이 흥미를 느낀다는 '자기 생각만의 과학'이란 무엇이고 교육적으로 어떤 의미를 지니는지 탐구하는 것은 인지적・정의적 성취가 일치하지 않는 현상의 실마리를 제공할 것이다. 첫 번째 학생은 좋은 평가를 받고자 하는 동기라는 측면이 강한 경우였다면, 두 번째 학생은 좋아하는 과학과 학교 과학을 다르게 인식하고 시험을 보기 위한 학교 과학에 별다른 노력을 기울이지 않은 경우다. 이와 달리 과학을 좋아하지만 진로와 관계가 없어 과학을 멀리한다는 김준홍 학생의 답변도 주목할 만하다. 많은 학생들이 좋은 성적을 얻기 위해서 성취하려는 노력이 병행되어야 한다고 말했고 높은 성취를 위해 과학에 대한 정의적 특성이 도움은 되지만 반드시 긍정적일 필요는 없다는 점을 피력했다. 학생들은 두 영역 간 불일치 현상을 경험을 통해 자연스럽게 인식했다. 진로와 직업 선택을 눈앞에 둔 고등학생들이 과학을 공부하도록 하는 것은 과학을 향한 즐거움과 흥미보다는 무엇을 얻기 위한 도구적 동기일 수 있다. 그러나 과학 교육은 한 사람 안에서 인지적 영역과 정의적 특성을 긍정적으로 하려는 분명한 목적을 가지며(MOEST, 2011), 정의적 영역 안에서도 여러 구인들이 조화와 균형을 이루어야 한다. 따라서 즐겁지 않은 과학 공부를 열심히 하는 사례, 재미와 흥미는 있지만 공부하고 싶지 않은 사례, 재미있어서 공부해도 좋은 성적을 받지 못하는 등의 불일치 사례들에 주목할 필요가 있다.

2. 검사 도구 결과와 면담 간 불일치 현상

면담 대상 학생들에게 정의적 특성 검사 결과를 보여주지 않고 스스로 자신의 성취 수준을 평가해 보도록 했다. 가로축은 인지적 성취, 세로축은 정의적 성취를 나타내는 그래프에서 성취 수준을 각각 둘로 나눈 네 영역 중 자신이 어디에 위치하는 지 이야기했다. 33명 중 13명의 학생들만이 양적 결과 자료에 근거해 연구자가 분류한 유형과 일치하는 유형으로 스스로를 평가했다(Table 10). 예를 들어 나현욱 학생의 경우 학업 성취도는 4분위이지만 정의적 성취는 1분위로 두 영역의 성취가 크게 불일치한 B유형으로 분류된다. 그러나 학생은 스스로 성적이 낮지만 정의적 성취는 중간 정도라고 말하며 A 아니면 D라고 응답했다. 따라서 나현욱 학생은 인지적 성취를 낮게 말하고, 정의적 특성은 높게 보려는 경향이 있으므로 '중립형, 인지적 성취 하향, 정의적 성취 상향'으로 특징을 표기했다. '과목 분리형'은 과학을 한 교과로 인식하지 않고 과목마다 다르게 나타낸 경우를 나타내며 학생이 과목마다 인식한 성취 수준을 모두 적었다. 12명의 학생들은 인지적 성취를, 16명의 학생들은 정의적 성취 수준을 양적 자료와 달리 인식하여 응답했다. 과목 분리형은 5명, 중립형은 3명으로 나타나 과목 간 성취에 불일치를 인식하는 학생들을 확인할 수 있었다. 학업 성취도가 낮은 A와 D유형 학생 11명중 자신의 인지적 성취를 상향하는 학생은 1명에 불과한 반면, 성적이 우수한 B와 C유형 학생 22명 중 성적을 하향하는 학생은 11명이었다. 또한 정의적 성취가 우수한 C와 D유형 학생 19명 중에는 정의적 성취를 하향하는 학생이 5명인 반면, 정의적 성취가 낮은 A와 B유형 학생 14명 중 정의적 성취를 상향하는 학생은 11명이었다. 즉, 성적이 높은 학생은 자신의 성적을 낮추어 말하고, 정의적 성취가 낮은 학생은 검사 결과와 다르게 과학을 좋아한다고 말하는 불일치 경향을 확인할 수 있었다.

학업 성취도와 정의적 특성 검사에서 나타난 양적 결과와 학생이 직접 이야기한 질적 결과가 다르게 나타나는 현상에서 두 가지 시사점을 찾을 수 있다. 첫째, 자기보고식 검사 도구가 유효한지 점검할 필요가 있다. 자신의 인지적 성취를 인식하여 평가하는 것은 결국 효능감이나 자아 개념 등의 정의적 특성과 관련 있다. 학생이 스스로의 상태를 진단하고 그를 말로 표현하는 것은 그 영역이 인지적이든지 정의적이든지 정의적 성취의 문제에서 출발한다. 결국 양적, 질적 결과에서 불일치 문제는 정의적 특성 검사가 학생의 실제 인식을 잘 반영하고 있는지의 문제로 좁혀진다. 실제 검사 결과와 상반된 말을 하고 있는 학생, 둘 중 어느 쪽이 실체에 더 가까운지 심층 분석이 필요하지만, 적어도 일치하지 않은 이유를 점검하는 과정은 정의적

특성 검사 도구를 정교화하는 출발점이다.

둘째, 성취 평가에서 드러나는 불일치가 학생들이 경험하는 학습과 평가 간 차이에서 오는 문제인지 확인할 필요가 있다. 평가는 학습 경험을 통해 나타나는 성취를 확인하는 과정인 동시에 학습 경험의 일부다. 교육과정이 학생들에게 제공하는 학습의 내용과 방법은 같지만 그 속에서 학생들이 겪는 학습 경험은 그들의 개인적 특성이나 물리적, 심리적, 사회적 상황에 따라 다를 수 있다. 다양한 성향과 상황을 지닌 학생들이 겪는 평가가 그들의 학습 경험 안에서 어떤 의미로 다가오는지 우선 파악해야 한다. 대체로 평가 도구로 특성이나 구인을 측정할 때, 학생들의 다양성은 어느 정도 통제된다는 가정하에 검사 결과를 해석한다. 평가 도구를 개발하는 연구자는 그 다양성에 의해 영향을 받거나 흔들리지 않는 도구를 제작하려는 반면, 다른 연구자는 여러 교육 맥락에 따라 검사 결과를 다방면으로 해석한다. 결과적으로 평가 도구는 측정 결과 자체만으로 객관적으로 해석되어야 하는 측면과 맥락에 따라 주관적 해석이 가능한 측면을 모두 지닌다. 두 측면은 서로 다른 전제에서 시작하는데, 전자는 다양성의 통제를 가정해야 하고 후자는 다양성에 따라 검사 결과가 달라질 수 있음을 인정한다. 간편하고 실용적 해석이 목적이라면 검사 과정에서 다양성이 발현될 수 있는 요인을 최대한 줄일 수 있도록 구체화하거나 정교화해야 한다. 맥락 변인에 따른 풍부한 해석에 목적이

Table 10. Comparison between students' test results and their self-evaluation

학생*	성별	성취도							본인 인식 유형**	특징
		인지적		정의적						
		T점수	분위	효능감	내적 동기	가치	평균	분위		
이경진	남	29.7	1	1.0	1.0	1.6	1.4	1	A	
조인진	남	32.5	1	3.0	1.4	1.7	1.9	1	D	정의적 성취 상향
서민영	여	34.9	1	3.3	3.0	3.3	3.2	4	C, D	중립형, 인지적 성취 상향
백예분	여	35.4	1	1.5	2.6	1.7	1.8	1	D	정의적 성취 상향
정선유	남	35.6	1	3.0	2.6	3.5	3.2	4	D	
지은영	여	35.8	1	1.7	2	1.9	1.9	1	A	
박기원	남	35.9	1	2.2	3.0	3.3	2.9	3	A, D	과목 분리형, 정의적 성취 하향
김준홍	남	39.3	2	2.8	2.6	3.2	3.0	3	D	
김인정	남	41.6	2	3.5	4.0	3.9	3.8	4	D	
박홍일	남	42.4	2	1.0	1.0	1.7	1.4	1	D	정의적 성취 상향
임주미	여	43.6	2	2.3	2.4	2.2	2.3	1	D	정의적 성취 상향
주경진	여	48.4	3	2.5	2.8	2.5	2.5	2	C, D	과목 분리형, 정의적 성취 상향, 인지적 성취 하향
김승주	남	49.1	3	2.5	2.2	3.0	2.7	2	A, C, D	과목 분리형, 정의적 성취 상향, 인지적 성취 하향
나현욱	남	62.4	4	1.5	1.6	1.3	1.4	1	A, D	중립형, 인지적 성취 하향, 정의적 성취 상향
안정현	남	62.3	4	2.3	2.8	2.9	2.8	2	B	
이가은	여	60.4	4	2.5	2.8	2.9	2.7	2	D	인지적 성취 하향, 정의적 성취 상향
나혜주	여	64.1	4	2.2	2.8	2.9	2.7	2	A, C	과목 분리형, 인지적 성취 하향, 정의적 성취 상향
윤진서	여	60.5	4	2.7	2.6	2.7	2.7	2	D	인지적 성취 하향, 정의적 성취 상향
강사덕	남	60.4	4	2.5	2.6	2.5	2.5	2	D	인지적 성취 하향, 정의적 성취 상향
이가현	여	62.6	4	2.8	2.6	3.1	2.9	3	B	정의적 성취 하향
김연수	여	61.5	4	2.7	2.8	2.9	2.8	3	D	인지적 성취 하향
조인석	남	63.9	4	4.0	3.4	3.8	3.8	4	B, C	중립형, 정의적 성취 하향
김민수	남	64.3	4	3.3	3.8	3.5	3.5	4	C	
현민조	남	65.7	4	4.0	3.8	3.7	3.8	4	C	
장지훈	남	64.6	4	3.0	3.2	4.0	3.6	4	C	
이환익	남	65.9	4	4.0	4.0	3.7	3.9	4	D	인지적 성취 하향
김규서	남	65.7	4	3.7	3.2	3.6	3.5	4	C	
박지현	여	63.3	4	4.0	4.0	4.0	4.0	4	C	
김별이	여	63.4	4	3.2	4.0	4.0	3.8	4	C	
박정현	여	62.7	4	3.2	3.6	3.9	3.7	4	B, C, D	과목 분리형, 인지적 성취 하향, 정의적 성취 하향
최상진	남	63.7	4	3.7	3.6	3.7	3.7	4	C	
김수현	남	65.0	4	3.7	3.8	3.8	3.8	4	B	정의적 성취 하향
양갑군	남	64.7	4	4.0	4.0	3.8	3.9	4	D	인지적 성취 하향

* 제시된 학생의 이름은 모두 가명임.

** A(C): 인지적 성취와 정의적 성취 모두 낮음(높음), B: 인지적 성취 높고, 정의적 성취 낮음, D: 인지적 성취 낮고, 정의적 성취 높음.

있다면 다양한 상황을 아우를 수 있는 명확한 평가 상황을 구현해야 하는 과제가 남는다. 이러한 결과 해석의 기능을 우선 염두에 두지 않으면 학생들의 다양한 특성과 학습 경험이 반영된 결과를 일률적으로 해석하거나, 의미 있는 분석이 어려운 제한적 결과를 유도하는 평가 도구를 제작하게 된다. 따라서 어떠한 기능을 선택하더라도 학생들의 개인적 특성과 학습 경험을 통해 드러난 다양성을 이해하여 평가 도구에 반영하려는 시도는 필요하다.

3. 검사 도구 결과와 면담 간 불일치 현상의 원인과 유형

가. 학생들의 생각에서 드러난 불일치 현상의 원인

과학 교육 평가 현장에서 양적으로 측정한 성취도와 학생들이 스스로 인식하는 성취의 결과가 일치하지 않는 원인을 검사 도구의 측면에서 살펴보았다. 첫 번째 원인은 검사 도구 자체가 지닌 측면으로서 측정하려는 개념의 모호성, 하위 구인들 간 불일치, 측정 시점과 척도의 구성 등이고, 두 번째는 검사 상황에서 학생들이 나타내는 반응 왜곡으로 사회적 바람직성과 자아 방어 기제 등이 관련된다.

1) 검사 도구와 학생들의 상호작용

가) 측정 대상 개념의 모호성: 과학과 학교 과학

Krynowsky(1988)는 과학 교육에서 학생의 태도를 측정하는 데 있어 개념적 명료성의 부족을 지적했다. 특히 과학이라는 용어가 여러 의미로 사용됨으로써 과학에 대한 태도가 무엇에 대한 태도를 측정하려는 것인지 혼동이 있을 수 있다고 주장했다. 보통 검사 문항 개발 과정에서는 학생들의 발성 사고 면담을 통해 모호한 용어를 다듬고 수정하기도 한다(Abd-El-Khalick *et al.*, 2015; Fives *et al.*, 2014). 학생들은 검사 상황에서 '과학'이 무엇을 의미하는지 질문했다. 네 과목으로 분리된 것 중 어떤 것을 염두에 두고 반응해야 하는지 묻기도 하고, 과학이 학교 과학인지 '그냥 과학'인지를 질문했다.

> 과목을 두고 어떤 과목을 생각하면서 채점(반응)을 했느냐가 은연중에 반영되지 않았나, 과학이 정말 제가 극과 극인 상황이어서, 정말 그 중에서 관심 없는 분야가 나오면, 아 하지말까 아 안 하고 싶다 이런 느낌이라서 이걸 할 때 문득 문득 어떤 과목이 생각이 들었냐가 반영된 거 같아요.
>
> (B유형 학생 이가은 면담 중)

> 저는 과목마다 되게 달라가지고, 약간 물리나 생명(과학)은 제가 생각하는 과학이랑 좀 달라요. 물리는 수학 같고 생명(과학)은 너무 암기 과목 같고 그래서 되게 좀 어렵고 해서 좀 싫다고 (반응)했죠.
>
> (B유형 학생 김승주 면담 중)

이가은과 김승주 학생은 정의적 성취가 학업 성취에 비해 낮은 B유형에 속했지만 자신들은 과학을 좋아하고 흥미가 높다고 말했다. Table 11은 선호 과목에 따른 인지적 성취도와 정의적 성취도의 평균을 분산 분석과 공변량 분석 결과로 나타낸 것이다. 특정 과목을 좋아하는 집단에 따라 학업 성취($F=9.56$, $p<.01$)와 정의적 성취($F=6.88$, $p<.01$)의 차이가 모두 통계적으로 의미 있게 나타났다. 다만 정의적 성취가 인지적 성취와 상관이 있어 인지적 성취가 설명하는 분산을 제거하기 위해 인지적 성취도의 T점수를 공변수로 하여 공변량 분석을 실시했다. 공변수인 학업 성취도의 T점수에 대한 선호 집단 간 차이는 통계적으로 유의미해서($F=26.99$, $p<.01$), T점수를 공변수로 선정한 것이 타당했다. 인지적 성취의 영향력을 통제한 후 교정된 정의적 성취도의 집단 간 차이도 역시 유의미해서($F=3.15$, $p<.05$), 특정 과목을 좋아하느냐에 따라 정의적 성취가 다름을 확인할 수 있었다. 설문 문항에서 좋아하거나 싫어하는 과목을 떠올림에 따라 반응이 달라질 수 있다는 점은 자연스럽게 추론할 수 있는 결과지만, 선호하는 과목에 따라 반응의 경향이 다르다는 점 또한 양적 자료와 학생들의 실제 반응이 다른 점을 뒷받침할 수 있다.

평가 도구에서 개념적 모호성은 과학과 과학 과목이 설문 문항에서 혼용되면서 드러났다. 실제로 학생들은 일반 과학과 학교 과학을 분리해서 생각하는 경우가 있었는데 검사지에서는 이 두 용어를 함께 썼다. 문항을 유심히 읽지 않는 학생들은 일괄적으로 과학 과목 혹은 과학으로 생각하는 경우, 과학과 과학 과목을 구별하는 경우, 편의적으로 해석하여 반응하는 경우 등이 있었다.

> 교과에 대한 흥미는 없지만 과학에 대한, 이게 되게 다르다고 생각을 해요, 저는 과학 교과목의 과학이랑, 그냥 제가 생각하는 좋아하는 과학은 달라요.
>
> (A유형 학생 백예분 면담 중)

> 학교에서 내는 설문지니까 과목이라고 생각하고 풀었겠죠. 그런데 여기는 위에 한번 과목이 나와 있었는데 언급이 안 되어 있기에 그냥 마음대로 해석해서 했어요.
>
> (D유형 학생 정선유 면담 중)

Table 11. Cognitive and affective achievement according to favorite subject

		성취도 평균								
		인지적(4과목의 T점수 평균)			정의적					
								성취도 T점수에 대한 공변량 분석		
		평균	표준편차	F	평균	표준편차	F	평균	표준오차	F교정 값
선호 과목	물리(N=32)	54.5	8.1	9.56**	3.1	0.54	6.88**	3.0	0.09	3.15* ($F_{T점수}$= 26.99**)
	화학(N=56)	54.5	8.1		3.0	0.54		3.0	0.07	
	생명과학(N=140)	48.9	8.8		2.8	0.46		2.8	0.04	
	지구과학(N=80)	48.2	9.1		2.7	0.57		2.7	0.06	
계(N=308)		50.3	9.1		2.8	0.53		2.9	0.03	

*P<.05, **p<.01

Table 12. ‘science’ words in test of science-related affective area

		과학 관련 정의적 특성 검사	
		해당 문항 번호	과학 관련 용어 사용 예(문항 번호)
자아 개념		1~6	과학 과목(1~6)
내적 동기(즐거움)		7~11	과학 지식(7), 과학을 배우는 것(8), 과학 공부(9), 과학에 관한 책(10), 과학 문제(11)
가치	일반적	12~16	과학과 기술(12, 14, 16), 과학(13, 15)
	개인적	17~21	과학(17, 18, 20, 21), 과학 개념(19)
	실용적(도구적)	22~26	과학 과목(22~26)

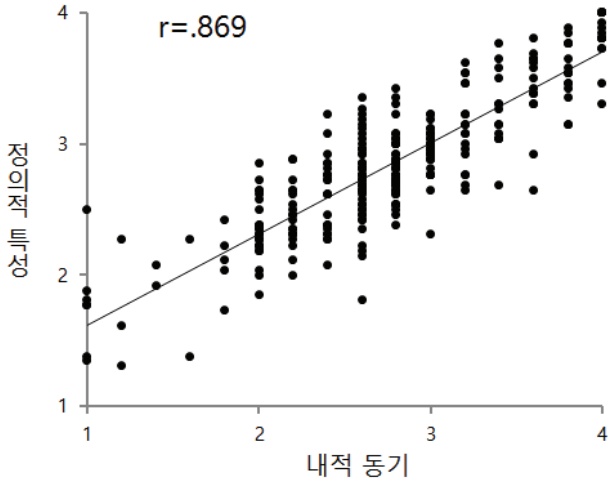

Figure 2. Mean of affective scales to intrinsic motivation score

> 일단 좀 자기가 생각하는 게 다른 거 같아요. 그니까 나는 (과학에) 관심은 있지만 이렇게 많이 흥미도 있고 관심도 있지만, 이런 학교 수업은 관심이 없다. 여기서는 흥미가 없다고 막 했잖아요. 그렇지 않다 이렇게. 이런 건 아마, 이건 제가 봤을 때 과목 위주로 한 것 같아요. 과목 생각으로.
> (A유형 학생 조인진 면담 중)

백예분은 인지적 성취와 정의적 성취가 낮은 A유형의 학생이었지만, 면담 과정에서 자신이 얼마나 과학에 대한 호기심이 높고 흥미 있어 하는지 적극적으로 말했으며 자신이 생각하는 과학과 학교 과학은 다르다고 강조해서 언급했다. 정선유는 과학 과목으로 생각하다가 편할 대로 반응한 경우, 조인진은 학교 수업과 연관된 과학이 싫어서 낮게 반응했지만, 실제로 일상의 과학은 좋아한다고 말했다. Table 12는 PISA 2006에서 발췌한 과학에 대한 정의적 특성 문항에서 사용한 ‘과학’ 용어를 정리한 것이다. 자아 개념과 실용적 가치 항목에서는 명시적으로 과학 과목이라는 용어를 사용했지만 다른 구인에서는 과학이라는 용어만 단독으로 사용하거나 그에 연결하여 개념, 기술, 공부, 문제 등의 단어를 썼다. 과학 과목을 명시한 것은 학습과 관련한 인식을 알아보기 위한 것이지만 학생들은 학교 안에서의 과학 학습과 과학에서의 실제 연구, 학교 밖 과학 등 다양한 상황을 떠올렸고 임의적으로 생각하고 반응한 경우도 나타났다. Kind, Jones, & Barmby(2007)는 태도 개념의 명확성 부족을 지적하며 학교 밖 과학과 실제 연구에 대한 태도 등을 구별하여 태도 측정 도구를 개발했다. Ku *et al.*(2016b)은 PISA 2018 예비 검사를 시행하기 위한 기반 구축 연구에서 설문 문항을 제작하는 과정을 공개했다. 내용 전문가, 교사, 평가 전문가, 문항 개발 전문가 등 다양한 집단이 개발 과정에 참여하지만, 실제 학생들의 인식을 알아보는 단계를 고려하여 특히 정의적 영역에서 측정하려는 구인을 보다 명확히 정의할 필요가 있다. Osborne *et al.*(2009)은 문항 제작과 선별 과정에서 전문가 집단을 과신하는 경향이 있으나 참여자와 반응이 같지 않음을 지적하며 참여자의 면담을 포함할 것을 제안했다.

나) 하위 구인들 간의 불일치

Figure 2는 과학에 대한 즐거움과 흥미를 나타내는 내적 동기의 점수를 모든 구인을 포함한 정의적 특성의 평균 점수와 함께 나타낸 것으로, 가로축과 세로축은 리커트 척도 1~4점을 나타낸다. 둘의 상관 계수는 .869($p<.01$)로 높지만, 내적 동기 점수에 대한 정의적 특성 평균은 다소 흩어진 분포를 보인다. 이는 전반적 정의적 성취에 비해 내적 동기가 상대적으로 높거나 낮은 정도를 나타낸다. 추세선 아래쪽에 분포하는 경우는 정의적 성취보다 과학을 즐겁게 느끼는 정도가 더 높은 것이고, 추세선 위쪽은 전반적 정의적 성취에 비해 과학을 좋아하고 즐거워하는 정도는 낮은 경우다. 정의적 특성 검사는 3가지 구인인 자아 개념, 내적 동기, 가치로 구성되었다. 이들이 하나의 차원으로 묶여서 하나의 척도로 사용될 수 있는지 여부는 요인분석으로 알아낼 수 있다(Kind, Jones, & Barmby, 2007). 면담에서 학생들은 과학에 대한 전반적인 느낌이나 흥미의 정도를 말하도록 요청받았다. 정의적 성취를 스스로 진단해 보는 과정이었고 학생들은 즐거움이나 흥미의 관점에서 자신의 위치를 판단했다. 그러나 검사지의 구인은 내적 동기 외에 자아 개념과 가치 등을 포함하고 있어서 내적 동기가 다른 구인에 비해 현저히 높거나 낮은 경우에는 검사에 의한 결과와 면담 결과가 일치하지 않는 원인이 될 수 있다.

Table 13. Tendency of response

구인	대표 진술	일치 반응 학생 수(%)			왜곡 반응 학생 수(%)					
					하향 반응			상향 반응		
		남	여	합계	남	여	합계	남	여	합계
자아 개념	잘 한다	100(32.5)	87(28.2)	187(60.7)	60(19.5)	42(13.6)	102(33.1)	17(5.5)	2(0.6)	19(6.2)
	못 한다	150(48.7)	121(39.3)	271(88.0)	10(3.2)	8(2.6)	18(5.8)	17(5.5)	2(0.6)	19(6.2)
내적 동기	좋다	154(50.0)	118(38.3)	272(88.3)	10(3.2)	4(1.3)	14(4.5)	13(4.2)	9(2.9)	22(7.1)
	싫다	151(49.0)	109(35.4)	260(84.4)	10(3.2)	5(1.6)	15(4.9)	16(5.2)	17(5.5)	33(10.7)
가치	중요하다	147(47.7)	119(38.6)	266(86.4)	7(2.3)	2(0.6)	9(2.9)	23(7.5)	10(3.2)	33(10.7)
	중요하지 않다	148(48.1)	116(37.7)	264(85.7)	9(2.9)	3(1.0)	12(3.9)	20(6.5)	12(3.9)	32(10.4)

다) 측정 시점과 척도의 구성

그 밖에 평가 도구와 학생들이 상호작용하는 과정에서 발생한 불일치 원인으로 분석된 것은 양적 자료의 측정 시점과 면담 시점의 간격과 척도의 구성 측면이었다. 학생들이 과학의 인지적・정의적 성취에 대하여 스스로 진단할 때 이환익 학생과 김연수 학생처럼 대체로 최근 시험 성적의 영향을 이야기하는 경우가 있었다. 특히 인지적 성취를 하향하는 학생들은 평가 결과에 따라 자신에 대한 인식이 달라졌다. 따라서 검사 도구에는 가변적 상황 또는 전반적 상황 여부를 명확히 지시하거나 검사 상황에서 이를 통제하려는 노력이 필요할 것이다.

> 아무래도 이번 성적이 좀 안 좋아서, 1학기 성적은 그냥 하는 대로 나왔으니까, 아 그럼 착각한 거네요.
>
> (C유형 학생 이환익 면담 중)

> 그 때는 제가 이번 시험을 보기 전이라, 이번 시험을 보고 나서는 생각이 바뀌었죠. 특히 과학이 많이 떨어졌어요.
>
> (C유형 학생 김연수 면담 중)

척도의 구성 측면에서 4단계 리커트 척도에서 반응의 강도가 높지도 낮지도 않은 중간 단계라고 생각하나 척도에는 중간 단계가 없어서 임의 반응했다고 말하는 학생들도 있었다.[2)]

2) 검사 상황에서 학생들의 인지와 심리 상태

가) 반응 왜곡: 사회적 바람직성

Messick(1991)은 자기보고식 심리 검사에서 나타나는 반응 왜곡을 반응 양식과 반응 세트로 구분했다. 반응 양식은 검사에 상관없이 비교적 일반성을 보이는 개인의 특성인 반면, 반응 세트는 검사 상황에 따라 발생하는 현상이다. 참여자의 주관적 경험 보고에 의존하는 자기보고식 검사는 반응의 진실성 여부가 검사 결과의 타당도와 신뢰도를 위협하는 한계로 작용하기도 한다(Bae, Lee, & Ham, 2015). 정의적 특성 검사 상황에서 반응 양식과 반응 세트 측면에서 자신의 솔직한 반응에 반해 의식, 무의식적으로 다른 반응을 선택한 학생들이 발견되었다.

Table 13은 세 가지 구인에 대한 대표 진술을 중심으로 학생들이 생각하는 자신의 반응 경향을 일치, 하향, 상향으로 나타낸 것이다. Table 2에서 반응 경향을 선택형으로 질문한 37번~42번에 해당하는 6문항에 대한 응답 결과이다. 설문에 응하는 자신의 느낌과 비슷하게 반응하는 사람을 찾도록 지시한 후 3인칭인 세 사람을 선택지에 등장시켜 답을 하도록 했다. 예를 들어 '민수는 자신이 과학을 잘 한다고 생각하지만, 한 단계 높여서 매우 잘 한다고 표시한다.'와 같은 식이다. 민수를 고른 학생은 원래 능력보다 상향하려는 왜곡 반응 경향이 있는 학생이다. 한편, 다른 선택지에는 한 단계 낮추어 반응한다는 하향 반응을 하는 학생과 솔직히 자신의 생각대로 반응한다는 일치 반응 학생을 구별한다. Table 13에서 6개의 대표 진술은 자기보고식으로 측정되는 정의적 영역의 하위 구인인 자아 개념(잘 한다/못 한다), 내적 동기(좋다/싫다), 가치(중요하다/중요하지 않다)를 나타낸다. 설문에 응하는 학생을 1인칭으로 설정하지 않은 이유는 구인에 대한 학생의 의견을 알기 위한 것이 아니라 가상적 상황에 대한 반응을 알아보려는 의도에서다. 자신이 생각한 대로 반응한다는 일치 반응은 '잘 한다고 생각하면 잘 한다'에 솔직히 반응한다는 경우가 60.7%로 가장 낮게 나타나 '잘 한다고 생각해도 못 한다(하향 반응)'(33.1%)고 하거나 '더 잘 한다(상향 반응)'(6.2%)고 왜곡하는 학생이 상당수 발견되었다. 특히 자아 개념에서 상향 왜곡 반응하는 남학생 수는 17명(5.5%)으로 2명(0.6%)인 여학생에 비해 더 잘하거나 덜 못하는 것으로 반응하려는 경향이 나타났다. 반면, 자아 개념의 '못 한다'에 대한 반응이나 다른 구인인 내적 동기와 가치 항목에서는 자신이 생각하는 바를 솔직하게 쓰지 않고 하향하거나 상향한다는 반응이 11.7~15.6%로 비교적 고르게 나타났다. 주목할 점은 학문적 능력에 대한 확신인 자아 개념을 제외하고 좋고 싫음의 선호를 나타내는 내적 동기와 과학을 중요하게 생각하는 가치 영역에서 모두 상향하려는 반응 경향이 있었다. 좋으면 더 좋고(7.1%), 싫어도 덜 싫다고(10.7%) 하며, 중요하다고 생각하면 더 중요하고(10.7%) 중요하지 않다고 생각하더라도 중요하다(10.4%)에 표시했다. 자신의 능력을 저평가하면서, 과학을 본심보다 더 좋아하고 더 중요하다고 반응하는 학생들이 일정 비율 존재하는 것이다. 이는 자기보고식 검사 결과를 드러난 그대로 해석할 것인지에 대한 중요한 판단을 요구하는 부분이다. 이러한 현상이 반응 양식으로서 개인이 지니는 일반적 특성인지, 반응 세트로서 검사 상황에 따라 발생하는 현상인지 연구자마다 다양한 의견을 제시한다(Au, 2007; Stephen, 2000; Paulhus, 1998; Ones, Viswesvara, & Reiss, 1996; MaCrae & Costa, 1983).

Crowne & Marlowe(1960)가 문화적으로 승인받으려는 방향으로 응답하려는 행동 양식을 사회적 바람직성(social desirability)으로 정

2) 본 연구에서는 중립 범주를 제공하는 경우 불성실한 반응을 유도할 가능성이 있다고 판단해서 반응 단계수를 4단계로 설정함.

Table 14. Achievement average by response groups

		'잘 한다'에 대한 반응 집단별 평균			집단별 분산 분석		
		일치(N=187)	하향(N=102)	상향(N=19)	F	p	사후 분석(Scheffe)
인지적 영역	4과목 평균 T점수	50.7	50.8	43.8	5.403*	.005	(일치+하향) vs 상향*
정의적 영역	자아 개념	2.7	2.5	3.0	5.625*	.004	일치 vs 하향* 상향 vs 하향*
	내적 동기	2.8	2.7	2.9	1.181	.308	-
	가치	3.0	2.9	3.0	2.495	.084	-
	정의적 영역 전체	2.9	2.7	3.0	3.330*	.037	-

*p<.05

의한 이후, Paulhus(1984)는 자기 기만적 고양(self-deceptive enhancement)과 인상 관리(impression management)라는 두 차원으로 사회적 바람직성을 구체화했다. 무의식적으로 자신을 긍정적으로 왜곡하려는 자기 기만적 고양과 의식적으로 좋은 이미지를 만들어 내려는 인상 관리는 반응 왜곡의 대표적 사례다. 특히 일종의 거짓말인 인상 관리는 상황이나 순간적 동기에 따라 그 정도가 달라진다(Ferrari, Bristow, & Cowman, 2005)는 점에서 검사 상황과 관련 있는 반응 세트에 해당한다. 반면 Paulhus(1998)는 자기 기만적 고양을 검사 상황에 따른 일시적 현상이 아니라 성격의 핵심적 측면으로 보았다는 점에서 반응 양식에 가깝다고 본다. 본 연구의 정의적 특성 검사에서 학생들은 자신의 솔직한 생각과 다르게 반응했다. 따라서 반응 왜곡, 특히 사회적 바람직성이 반영되었다.

> 뭔가 너무 못해 가지고, 체크하다 보니까 너무 못해서 그냥 하나씩 올렸어요. 익명이면 그냥 그대로 할 텐데.
>
> (D유형 학생 정선유 면담 중)

보다 긍정적인 모습으로 보이고 싶었던 정선유 학생은 익명이 아니어서 한 단계씩 올려서 반응했고 이는 인상 관리의 전형적 예다. 의식적이든 무의식적이든 자신을 좋은 사람으로 보이고 싶은 사회적 바람직성이 나타난다면 단계를 높여서 반응하는 상향 반응이 나타나는 것이 자연스러우며 Table 13에서도 이런 결과를 확인할 수 있다. 그러나 자신의 학문적 능력에 대한 확신은 오히려 깎아내리는 것이 더 자연스럽고 마음 편하다는 학생들도 상당수 발견되었다.

> 같이 얘기하는 상대방을 좀 추켜 세워주는 게 있어요. 저는 저를 막 좀 깎아 내리고요.
>
> (C유형 학생 김별이 면담 중)

> 매우 잘 한다고 이게 나 혼자만 생각 하는 건 괜찮은데 다른 사람하고 같이 다 보는 거면, 매우 잘한다 하면 뭐지? 약간 좀, 못하면, 다른 거 못하면 좀 압박감 같은 것도 들 수 있고. 그런 것 때문에 이건(매우 잘한다고 솔직하게 표시하는 것) 너무 깝치는(깝죽거리는) 거잖아요.
>
> (B유형 학생 강사덕 면담 중)

김별이와 강사덕 학생은 학업 성취도가 높았지만 다른 사람들의 시선을 의식하며 능력에 대한 반응을 한 단계씩 낮추고 있다고 했다. 이는 의식적으로 자신의 인상을 관리하는 측면이기는 하나, 자신을 보다 긍정적으로 꾸미기 위하여 오히려 낮추어 반응하는 겸손의 모습으로 나타났다. 따라서 정의적 특성 검사 도구의 자아 개념 구인에서 사회적 바람직성은 하향 반응을 유도함으로써 점수를 낮출 가능성이 있다. Martin, Mullis, & Foy(2012)는 TIMSS 2011에서 동아시아 국가 학생들의 높은 성취와 낮은 태도 영역 점수에 대하여 동양인의 겸양의 미덕, 학습에 대한 진지한 태도를 중시하는 문화 풍토와 높은 수준의 학습 내용 등을 이유로 추측했다. 그 밖에 아시아 국가는 서양 국가에 비해 사회적 바람직성이 높다는 문화 비교 측면의 연구(Dudley *et al.*, 2005; Keillor, Owens, & Pettijohn, 2001), 개인보다는 집단을 우선하는 한국의 문화 풍토가 개인의 태도에 미치는 영향(Bae, Lee, & Ham, 2015), 다양한 사회·문화적 배경과 언어 차이를 고려하지 못한 국제 비교 평가의 문제점(Chi, 2011; Murayama *et al.*, 2009; Suzuki & Ponterotto, 2007) 등을 지적한 연구들이 발표되었다. 이들은 우리나라에서 독특하게 나타나는 사회적 바람직성이 평가와 어떻게 연결되는지를 우선 탐색해야 한다는 시사점을 제시한다. 사회에서 환영하는 모습으로 자신을 보여주고자 하는 사회적 바람직성은 우리나라 사회 문화 속에서 변형되어 학생들에게 영향을 미친다. 이를 고려하여 정의적 영역 검사 결과를 해석하는 데 어떤 주의를 요하는지 밝히는 것은 과학 교육 평가에서 새로운 분야로서 타당하고 신뢰할 수 있는 평가 방안을 마련하는 근간이 될 수 있다. 특히 다른 나라와 차별화된 현재 우리나라 교육 현장에서 발견되는 학생들의 사회적 바람직성의 모습을 살피는 것은 그 시작이다.

Table 14는 자아 개념의 대표 진술인 '과학을 잘 한다'에 대한 반응 집단별 인지적·정의적 영역 평균을 분산 분석 결과와 함께 나타낸 것이다. 자신의 생각대로 솔직히 반응하는 일치 집단(50.7)과 낮추어 반응하는 하향 왜곡 집단(50.8)은 인지적 영역에서 동일한 수준의 집단이다. 반면 더 잘한다고 높여서 반응하는 상향 집단의 학업 성적(43.8)은 다른 집단보다 유의미하게 낮았다. 학업 성적이 높은 학생보다 낮은 학생들은 더 잘 한다고 상향하는 경향이 일부 있으나, 대부분의 학생들은 솔직한 반응과 낮추려는 반응을 하는 학생 간에 성적 차이가 거의 없는 동일 집단이다. 정의적 영역에서는 하향하는 학생의 평균(2.7)이 일치(2.9)하거나 상향(3.0)하는 학생들에 비해 유의미하게 낮았지만($F=3.330$, $p<.05$) 사후 분석에서 구체적 집단 간 차이가 나타나지 않았다. '잘 한다'의 진술은 정의적 영역의 세부 구인인 자아 개념의 대표 진술이므로 이를 분산 분석한 결과 하향하는 집단의 평균(2.5)이 일치(2.7)하거나 상향(3.0)하는 집단에 비해 유의미하게 ($F=5.625$, $p<.05$) 낮았다. 이는 실제 성적에 차이가 없음에도 자신을 낮추는 학생들은 점수를 낮추어 반응하는 경향이 있으며 자신을 높이는 학생들은 정의적 영역 평가에서 높은 점수를 주는 경향이 있음을 나타낸다. 학생들의 반응 성향이 반응 결과에 어느 정도 반영되었다.

다음의 학생들은 객관적 기준에 의해 자신을 평가하기보다 다른

사람의 시선을 의식하거나, 끊임없이 주변 친구들과 비교하면서 자신을 다른 이보다 낮게 평가한다. 특히 나현욱 학생은 의식적 인상 관리를 넘어서 자신의 능력을 항상 낮추어 생각하여 개인적 성향으로 굳어진 상태였다. 나현욱 학생에게 사회적 바람직성은 무의식적으로 자신을 속이되, 사회가 원하는 '겸손'의 모습으로 나타난다. 개별성보다는 관계성을 지향하는 한국인의 특성(Choi, 2004; Cho, 2003)상 서구의 독립성을 중시하는 문화에서 개발된 사회적 바람직성 척도와 다른 양상이 나타날 수 있다. 우리 사회에서 바람직하게 생각하는 사람은 자기를 낮추어 타인을 높이는 겸양의 미덕을 발휘하는 사람이다. 따라서 사회적 바람직성이 높은 학생은 오히려 자신의 평가 점수를 낮춘다.

> 기준이 다 다르니까 뭔가 다른 사람이 보기에, 아니 저는 제가 잘한다고 생각할 수 있지만 다른 사람이 보기엔 아닐 수도 있으니까 다른 사람이 보는 눈에 맞춰서….
>
> (B유형 학생 윤진서 면담 중)

> 객관적으로 높다고 생각되는 등급 같은 지표가 몇 개 나와도 저는 그냥 자신이 좀 못하는 거 같다고 느껴요. 주변에 친구들 보면, 되게 잘하는데, 저만 좀 낮은 거 같아서요. 그냥 절대적으로는 꽤 높은 편이라고 생각은 해요. 그런데 비교를 하면 좀.
>
> (B유형 학생 나현욱 면담 중)

나) 자아 방어 기제 사용

한편 자신에 대한 억제(suppression)를 이유로 자신을 낮추는 반응을 한다고 말한 학생들은 일종의 자아 방어 기제를 사용한 것으로 보인다. 자신의 위치를 표시하라는 외부적 요구에 솔직하게 표시하고 싶은 마음과 이를 자만이라고 생각하는 마음이 갈등으로 작용한다. 그리고 자만하면 발전할 수 없다는 의식적 혹은 무의식적 자아의 판단은 한 단계 아래로 반응하는 행동(Vailant, 1992)으로 나타난다. 이는 개인의 적응과 성숙 수준을 나타낸다는 점에서 개인적 특성에 가까운 반응 양식에 해당하지만 결과적으로 평가 점수를 낮추는 결과로 이어졌다.

> 아 뭔가 나한테 한계를 두는 거 같다고 해야 되나? 뭔가 매우 잘한다고 생각을 하는데 왠지 난 이제 매우 잘하니까 더 안 해도 돼, 약간 이런 느낌? 나의 한계를 내가 정하는 느낌인 거 같아요.
>
> (B유형 학생 이가은 면담 중)

> 저는 완전 극악으로 가야지 깨달음이 있거든요. 그러니까 완전 너는 완전 싫어 그래야지 제가 발전이 되는 거 같아서 거기다가 표시를 한 건데. 그러니까 계속 나는 너무 잘 해라고 생각하게 되는. 매우 잘 한다고 계속 적게 되면 저 자신이 태만해지게 된다고 생각을 해서 그래도 한 단계라도 낮춰서 더 열심히 해야 되지 않을까라는 생각이 들어요.
>
> (C유형 학생 장지훈 면담 중)

결국, 검사에서 자신의 생각을 바로 드러내 반응하지 않는 학생들이 많았고, 솔직하게 반응했다는 내용에서도 의식적 혹은 무의식적 방어 기제가 작용하기도 했다. 따라서 자기보고식 평가에는 본심과 다른 반응이 존재함을 인정하고 결과의 해석에 이를 반영해야 할지 면밀히 검토할 필요가 있다. 평가의 목적이 개인의 특성 파악이나 선발에 있다면 이런 현상이 특정 학생에 국한된 개인적 성향인지, 검사 상황에서 학생들이 겪는 전반적 심리 상황인지 구분해야 검사의 실효성을 얻을 수 있다. 또한 교육 프로그램의 효과를 비교하고 평가하여 정책을 결정하는 것이 목적이라면 전반적 왜곡 반응이 불러올 파장은 더욱 커진다. 학생들의 다양한 인지, 심리, 정서 상태가 검사 도구와 상호작용하면서 내어놓는 결과는 측정하려던 것을 단순히 수치화해서 내어 놓는 간단한 결과가 아닐 수 있다는 점을 유념해야할 것이다.

나. 자기보고식 검사 도구를 사용하는 현장에서 나타난 학생들의 불일치 반응 유형

자기보고식 검사 도구를 사용하는 평가 현장에서 나타나는 불일치 현상은 측정한 결과 값을 해석할 때 유의해야 할 점에 대한 단서를 제공한다. 검사에서 학생들의 반응은 크게 두 유형으로 나타났는데 두 경우 모두 도구를 중심으로 발견되는 불일치 반응과 관련 있다. 첫째는 검사 도구 관련 불일치 반응, 둘째는 검사 상황 관련 학생들의 심리 상태로 인한 불일치 반응이다.

1) 검사 도구 관련 불일치 반응: 문항 개발자가 의도한 개념과 학생들이 이해하는 개념 간 차이

도구를 개발하는 연구자는 측정 개념이 학생들의 마음속에서도 의도한 그대로의 의미로 다가갈 수 있을지 사전에 살펴야 한다. 그렇지 않은 경우, 학생들은 검사 도구의 문항을 읽으면서 이 도구에서 사용한 개념이 정확히 무엇을 의미하는 것인지 판단해야 하는 혼란을 경험한다. 학생들은 일상에서 흔히 사용하는 용어에 대해 저마다의 경험에서 축적한 수많은 형상과 의미를 가지고 있다. '과학'이라는 용어는 과학 교육의 정의적 영역 평가에 수없이 등장하며 과학에 대한 효능감, 선호, 가치 판단 등의 질문에 활용되어 왔다. Krynowsky (1988)는 '과학'이라는 용어가 과학 교수, 과학 직업, 과학 그 자체, 과학자의 일, 구체적 과학 쟁점, 학교 과목 등을 의미하면서 학생 뿐 아니라 연구자들도 이를 혼동해서 쓰고 있다고 지적했다. 1990년대에 Costa(1995)는 일부 학생들에게 학교와 과학은 일상과 유리된 '다른 세계'로 인식된다고 주장했다. 2016년의 학생들에게 학교 과학은 여전히 다른 세계지만, 달라진 점이 있다면 일상 과학이 실용적이며 첨단을 달리고 있는데 비해 다음의 윤진서 학생의 경우처럼 학교 과학이 오히려 뒤떨어진 이론을 다루고 있다고 느낀다는 점이다.

> 학교 과학은 옛날 것까지 해야 하고 이론적이니까 뭔가 세상을 위해 할 수 있는 게 없는 거 같아요. 그러니까 적용하기가 쉽지 않은 거 같아요.
>
> (B유형 학생 윤진서 면담 중)

또한 학교 과학을 떠올릴 때는 자동적으로 시험이 연상되면서 어렵고 싫은 느낌을 갖게 되고, 이러한 기분은 일상 과학과 학교 과학을 분리해서 생각하게 한다.

> 과학은 실생활에서 쓰이고 더 실용적인 데서 많이 쓰이는데, 학교에서 배우는 과학들은 약간 내신이나 이런 시험 봐서 점수를 따기 위한 그런 목적이죠.
>
> (D유형 학생 김준홍 면담 중)

학생들은 과학에 대한 능력과 선호와 가치를 묻는 검사지 문항을 읽으면서 일상 과학인지, 학교 과학인지, 학교 과학이라면 어떤 과목인지 생각하면서 망설이는 경험을 한다. 아울러 용어의 명확성이 확보되더라도 학생들은 자신과의 관련성과 최근에 경험한 상황에 따라 판단을 다르게 하는 경향도 나타난다. 학교에서 배우는 과학 과목 중 자신이 좋아하거나 싫어하는 과목을 떠올리면서 반응하거나, 가장 최근에 일어난 인상 깊은 사건, 예를 들어 최근에 치룬 시험과 같은 경험을 통해 얻은 결과에 따라 반응의 방향과 정도를 달리했다. 앞서 면담했던 학생들(이가은, 이환익, 김연수)의 면담 내용이 그 예다. 학생들이 혼동할 만한 용어를 만나게 되었을 때의 반응은 두 가지로 나뉘는데, 용어의 의미를 적극적으로 질문하는 경우와 자의적으로 해석해서 반응하는 경우다. 대부분의 학생들은 용어가 혼란스럽다는 사실조차 인지하지 못한 채 검사를 마치기도 하므로 응답자의 혼란을 알아내는 것은 쉽지 않다. 따라서 문항의 의미를 적극적으로 질문하는 학생의 의견을 눈여겨 볼 필요가 있다. 더 나아가 검사 도구 개발 단계에서 이 과정을 반영하면 도구가 학생과 상호작용하는 과정에서 드러나는 용어의 모호성을 어느 정도 감소시킬 수 있을 것이다. 한편 학생들이 자의적으로 선택하려는 순간에 무엇을 떠올려 반응하는 지는 대체로 측정하려는 대상에 대한 전반적 느낌과 정서의 문제일 수 있지만 반대로 특정 상황과 시점의 문제일 수도 있다. 또는 문항마다 일관성 없는 기준이 적용될 수도 있다. 따라서 반응에 영향을 미치는 세부 요인을 현실적으로 파악하거나 그 결과의 의미를 해석하기 어렵다는 점에서 용어 혼란으로 야기되는 문항에 따른 임의 반응과 같은 불일치 현상을 최대한 막아야할 필요가 있다.

검사 도구와 학생의 상호작용에서 나타나는 불일치 반응의 감소 방안으로 첫째, 검사 도구 개발 단계에서 측정 개념을 명확히 하기 위해 이를 표현하는 용어의 의미를 구체적으로 한정하는 과정을 안내서로 만들어 활용한다. 안내서의 1단계는 문장을 이루는 단어를 우선적으로 검토하는데, 일상에서 자주 쓰는 용어일수록 여러 상황에 열린 의미를 지닐 수 있으므로 주요 단어의 일상적 의미를 함께 검토하며 개념을 분명히 한다. 단어 중심의 분절적 분석이 완료되면 2단계로 전체적 문장의 맥락을 살핀 후 마지막 단계로 의미를 명확히 하는 수식어나 조사, 관형사, 부사 등을 검토하는 단계로 진행할 수 있다. 둘째, 검사 대상자의 의견을 도구 개발 단계에 반영한다. 문항에 대한 참여자의 인지적, 정서적 반응을 살피고 모호한 문항에 대한 피드백을 얻는 과정에서 검사 도구와 학생 간 불일치 반응을 감소시킬 수 있다. 셋째, 검사 상황에 대한 구체적 안내를 포함하는 지시문을 활용한다. 지시문은 개념을 분명히 하고, 판단하는 시점이나 관점을 제한하는 동시에 검사 상황에 대한 풍부한 이해를 촉진할 수 있도록 구성되어야 한다. 검사 도구의 완성은 문항 양호도의 검증에 그치는 것이 아니라 그 문항에 대한 정확한 맥락을 부여하는 좋은 지시문의 완성을 반드시 포함해야 할 것이다.

2) 검사 상황 관련 학생들의 심리 상태로 인한 불일치 반응: 표현된 반응과 속마음 간의 차이

검사 도구에 얽힌 문제 이외에 검사 상황에서 학생들이 반응하는 양상은 반응 양식과 반응 세트에 의해 속마음과 다른 반응을 나타내는 불일치가 나타났다. 이러한 왜곡 반응은 잘 보이고 싶은 마음이 의식적으로 표현된 인상 관리와 솔직한 마음과 달리 자신을 포장하려는 반응을 무의식적으로 행하는 자기 기만적 고양이라는 형태로 나타난다. 이 두 가지 구인은 사회적 바람직성이라는 특성으로 묶이며 자기보고식 검사 도구로 측정한 결과에 대한 신뢰도와 타당도에 위협이 되는 대표적 오염원이다(Bae, Lee, & Ham, 2015). 과학에 대한 태도 문항에서 발췌한 자아 개념, 내적 동기, 가치 문항에 반응한 학생들은 학문적 능력에 대한 확신을 제외하고는 어느 정도 긍정적인 방향으로 점수를 높이려는 경향이 있었다(Table 13, Table 14). 그러나 과학을 잘 하는지의 능력에 대한 문항에서는 상당수 학생들이 점수를 오히려 낮추어서 이를 사회적 바람직성 특성으로 해석하기 어려운 경우도 나타났다. 사회적 바람직성이 사회적 지지를 받는 방향으로 응답하려는 경향성을 의미한다면, 원만한 대인 관계를 중시하는 한국 사회에서 겸손하게 능력을 낮추는 반응이야말로 사회적 바람직성의 또 다른 얼굴이라 할 수 있다. 또한 학생의 성숙 정도가 반영된 방어기제의 영향을 받아 반응하는 학생도 발견되었다. 이렇듯 학생들은 다양한 정서적, 심리적 요인들로 인해 반응의 정도를 낮추거나 높이는 등 속마음과 다른 반응을 하므로 그 결과 값의 해석은 다소 모호해질 수 있다.

지금까지 과학 교육 평가 영역에서는 이러한 불일치 현상의 상태와 원인을 점검하고 규명하려는 연구는 부족하다. 평가에서 불일치 현상을 파악하는 연구는 이를 보정하고 보완하는 평가 도구를 마련하는 연구로 진행될 뿐만 아니라, 학생들의 개인적 특성과 학습 경험에서 드러나는 다양성을 이해하게 한다. 과학 교육 이외 타 영역에서 관련 연구는 사회적 바람직성이 측정하려는 구인-성격, 동기, 사회적 규범 등에 미치는 영향을 인위적 조작 실험, 동기가 다른 집단의 비교, 자연스러운 검사 상황에서 진행해 왔으며(Kim, 2013; Son, Cha, & Kim, 2007), 문항에 척도를 투입하는 방법 뿐 아니라 통계적 방법으로 반응 적합도를 살펴보는 연구까지 다양하게 연구되었다. 특히 전화, 대면, 웹 기반, 익명, 실명 등의 검사 운영 모드에 따른 양상을 분석하거나 구직과 같은 선발 상황에서 사용할 수 있도록 보다 실용적 측면에서도 활발히 연구가 진행되고 있다(Au, 2007; Stephen, 2000). 그러므로 과학 교육 평가에서도 자기보고식 검사 도구를 활용하여 학습자와 교육 상황을 파악하고, 교육과정을 개선하며 나아가 선발에 연결되는 타당한 평가 도구를 얻고자 한다면 표현된 반응과 속마음의 간극을 줄이거나 적어도 그 차이를 탐지해서 해석할 수 있는 장치를 마련할 필요가 있다. 본 연구에서 드러난 학생들의 반응과 속마음과의 불일치 현상을 통해 검사 도구에서 나타난 결과 값을 의미 있는 해석으로 삼기 위해서는 이러한 불일치 반응에 주목하여 그 영향력을 파악하는 것이 선행 과제일 것이다.

Ⅳ. 결론 및 시사점

본 연구에서는 학교 현장에서 평가를 중심으로 한 양적 자료와 질적 자료에서 나타나는 몇 가지 불일치 사례를 양적, 질적 연구 방법을 활용해 분석했다. 인지적 성취와 정의적 성취, 검사 도구로 측정한 결과와 면담 결과 간 불일치 현상이 나타나는 이유를 검사 도구의 측면에서 고찰했다. 특히 서로 다른 특성을 지닌 학생들의 과학 학습 경험을 반영하고자 학생들을 인지적 성취와 정의적 성취 수준에 따라 네 유형으로 분류해서 유형별 다양한 경험과 인식을 비교했다. 그

결과 불일치 사례가 발생하는 일부 원인은 다양한 상황과 특성을 지닌 학습자가 평가 도구와 상호작용하는 과정에서 나타내는 반응이라는 측면에서 몇 가지 근거를 제시했다. 또한 학생들이 겪는 학습 경험에서 평가가 지니는 의미를 살펴보고 평가와 평가 도구에 대한 시사점을 도출했다. 이를 위해 인문계 고등학생 308명을 대상으로 과학 교과 4과목에 대한 학업 성취도 점수와 정의적 영역 특성 검사 결과를 양적 자료로 수집했다. 질적 자료로는 인지적・정의적 성취도를 네 유형으로 분류한 학생 33명의 면담 자료를 분석했다. 양적 자료와 질적 자료에서 나타나는 불일치 사례를 각각 분석한 후 두 자료 간 불일치 현상을 찾아냈고, 그 원인을 밝히기 위해 학생 면담 내용을 다시 분석하는 한편, 통계적으로 이를 입증할 수 있는 몇 가지 근거를 제시했다. 본 연구에서 발견된 결과를 요약하면 다음과 같다.

첫째, 인지적 성취도와 정의적 성취도의 양적 자료를 분석한 결과 두 영역 각각에서 과목별, 구인별 성취의 차이가 크게 나타나는 학생이 상당 수 발견되었고, 특히 인지적 성취도와 정의적 성취도 차이가 나는 학생은 20% 이상 나타났다. 인지적 성취도에서는 과학 교과의 4과목 간 성적의 상관이 높아서 대체로 과목 간 성취 경향이 일치했으나 선택적 학습이나 선호도의 영향으로 과목 간 성취 차이가 크게 나타나는 학생들이 있었다. 정의적 성취도에서는 하위 구인 간 상관은 비교적 높거나 높은 상관으로 나타났다. 인지적 성취도와 정의적 성취도의 상관은 비교적 낮아서 두 영역 간 성취 경향이 일치하지 않는 학생들이 상당수인 것으로 조사되었다. 이러한 불일치 현상은 선택한 진로와 진학을 위해 의도적으로 과학 학습을 조절하는 경우, 학교 과학과 과학에 대한 인식 차이 등의 사례로 나타났다.

둘째, 인지적 성취도와 정의적 성취도의 양적 자료와 학생들의 면담 내용을 비교한 결과 스스로 생산한 양적 자료와 다르게 자신을 평가하는 학생들이 대부분이었다. 학업 성취도가 우수한 학생은 자신의 성적을 낮추어 말하고 정의적 성취가 낮은 학생은 검사 결과와 다르게 과학을 좋아한다고 말하는 불일치 경향이 뚜렷하게 나타났다.

셋째, 과학 교육 평가에서 나타나는 두 가지 불일치 사례는 검사 도구와 상호작용하는 학생들의 다양한 특성에서 비롯된 반응으로부터 원인을 일부분 찾을 수 있다. 검사 도구와 관련 있는 불일치 유형은 '문항 개발자가 의도한 개념과 학생들이 이해하는 개념 간 차이'와 '표현된 반응과 속마음 간의 차이'로 나타났다. 검사 도구에서 사용한 용어가 학생들에게 모호하게 인식되는 경우 자의적으로 반응하거나 일관성 없이 응답했는데 특히 과학과 학교 과학에 대해 학생들이 지닌 상이한 생각을 반영한 반응들이 전자에 해당한다. 후자의 경우에서 자신의 학문적 능력에 대한 판단은 낮추려 하지만 다른 항목은 긍정적인 쪽으로 반응하려는 반응 왜곡이 감지되었다. 자기 기만적 고양과 인상 관리를 포함하는 사회적 바람직성뿐 아니라 자신을 억제하려는 자아 방어 기제도 사용되었다. 그 결과 학생들은 자신의 속마음에 비해 점수를 낮추거나 높여서 반응한다고 했고, 특히 점수를 낮추는 반응은 한국의 사회 문화적 맥락에서 총체적 분석을 요하는 독특한 부분이다.

이상의 학교 현장의 과학 교육 평가에서 나타나는 불일치 사례를 분석한 결과로부터 도출한 시사점은 다음과 같다.

첫째, 성취의 평가에서 나타나는 불일치 현상의 의미와 원인을 밝히려는 노력을 새로운 관점으로 시도해야 한다. 국제 학업 성취도 평가에서 발견되는 인지적 성취와 정의적 성취의 불일치 현상에 주목하고 그 원인과 실정을 파악하려는 후속 연구(Kim & Cho, 2013; Choe *et al.*, 2013; Kim *et al.*, 2009)가 진행되고 있다. 이들 연구의 궁극적인 목적은 정의적 특성의 함양이지만 새로운 관점을 보완해서 접근할 필요가 있다. 학습 현장에서 학생들이 지닌 다양한 인식과 학습 경험 간, 경험과 평가 간의 간극이 이들 불일치 현상의 핵심임을 인지한다면 이들을 심층적으로 파악하려는 시도와 함께 성취를 측정하는 도구 점검에도 관심을 가져야 한다.

둘째, 한 학생이 지니는 인지적・정의적 성취의 유형은 그 학생이 겪은 학습 경험의 결과를 진단할 뿐 아니라 미래 경험을 예측할 수 있는 예언적 기능을 지닌다는 점에서 유용한 지표다. 다양한 특성을 지닌 학생들에 대한 온전한 이해는 그들이 겪게 될 어려움을 사전에 예상하고 대비할 수 있도록 도움을 준다. 학생에 대한 온전한 이해는 인지적 성취와 정의적 성취가 조합하는 과정에서 나타나는 여러 현상을 설명하고 예상 가능케 함으로써 장차 학생이 겪게 될 경험의 반경을 미리 그려볼 수 있다. 따라서 온전한 이해를 촉진할 수 있는 도구의 개발과 함께 성취 유형에 따른 다양한 특성을 파악하는 연구도 보완되어야 한다.

셋째, 평가에서 나타나는 불일치 결과가 검사 도구에 내재한 약점에서 비롯된 것인지 점검해야 한다. 검사 도구의 양호도를 주로 통계적 방법에 의존할 경우 간과하거나 축소하는 본질적 현상이 있는지 수검자의 입장에서 우선 파악해야 한다. 도구가 학생들의 실제 인식을 반영하여 잘 작동하고 있는지 여부는 도구와 학생이 상호작용하는 중에 발생하는 불일치로부터 감지될 수 있다. 측정하려는 개념과 그에 따른 하위 구인에 대한 명확한 이해는 검사를 받는 대상 뿐 아니라 연구자 자신도 오해가 없는지 확인하고 성찰해야 한다. 도구에서 사용하는 용어가 학생의 경험과 일상 맥락에서 여러 의미를 갖지는 않는지, 전반적이거나 일시적 상황인지에 따라 응답이 달라질 것인지 등을 고려하는 것은 이러한 예에 해당된다. 그밖에 검사 도구에 내재한 약점 외에 검사 상황에서 학생들은 본심과 다른 왜곡 반응을 할 수 있음을 인지하고 결과를 어떻게 해석해야 할지 고민해야 한다.

넷째, 검사 도구 개발 단계에는 측정하려는 개념과 상황을 명확히 하려는 노력(Chung & Shin, 2016)이 반영되어야 한다. 이를 위해서는 측정 개념을 표현하는 용어의 의미를 구체적으로 한정하는 과정을 안내서로 만들어야 한다. 문항의 진술을 분절 단위에서 전체 맥락으로 분석하고 문장의 의미를 명확히 하는 부속 성분까지 검토하는 단계가 포함된다. 또한 전문가 집단 뿐 아니라 검사 대상자의 의견을 반영하는 단계, 검사 상황을 구체적으로 안내함으로써 문항에 대한 정확한 맥락을 부여하는 좋은 지시문 개발 단계가 포함되어야 한다.

본 연구에서는 과학 교육 평가에서 나타나는 불일치 사례에 대한 원인 중 하나로 평가 도구와 학생 간의 상호작용 중 발생하는 두 가지 유형의 불일치 사례를 중심으로 논의했다. 인지적・정의적 성취를 기준으로 분류한 유형에 따라 학생들의 학습 경험이 각기 다르며 이들이 평가를 어떻게 결부시키느냐에 따라 과학과 학교 과학에 대한 인식이 달라진다. 평가 도구가 과학과 학교 과학을 구분하지 못하면 학생들은 혼동된 상태로 검사에 반응하며 이는 불일치의 일부 원인이 될 수 있다. 그러나 인지적 성취와 정의적 성취가 서로 엇갈리는 불일치 현상은 원인을 오롯이 도구의 문제로만 돌릴 수 없는 현실이기도 하다. 학생들의 학습 경험 속에서 찾아낸 평가는 학교 과학에 수반된 부담감의 주요 원인이고, 주로 문제 풀이 형태로 제시되면서 이에

대비한 학습 경험을 획일적으로 고정시켰다. 결국 학교 과학에 대한 부정적 인식의 원인인 평가의 문제가 남는다. 국제 학업 성취도에서 우리나라 학생들의 낮은 정의적 성취를 우려하는 각계의 반응에 따라 다행히 정의적 특성을 함양하려는 다방면의 시도를 모색하고 있다. 그 출발점은 학습 경험의 일부이며 과정으로서의 측면을 지닌 평가가 제 역할을 다할 수 있도록 교육적 평가의 의미를 되살리는 것이다.

국문요약

이 연구는 과학 교육 현장에서 인지적・정의적 평가를 중심으로 한 양적 자료와 질적 자료에서 나타나는 몇 가지 불일치 사례를 분석했다. 308명의 고등학교 2학년 학생을 대상으로 학업 성취도와 정의적 성취도를 양적 자료로 수집했고, 그 중 33명의 학생을 면담한 질적 자료를 분석했다. 주로 검사 도구의 측면에서 불일치 사례의 원인과 유형을 고찰했다. 연구 결과 양적 자료인 인지적 성취와 정의적 성취 영역 각각에서 과목별, 구인별 차이가 크게 나타나는 학생들이 상당수 있었고, 특히 두 영역 간 성취도 경향이 일치하지 않는 학생들도 20% 이상 분석되었다. 선택한 진로와 진학을 위해 의도적으로 과학 학습을 조절한 사례, 학교 과학과 과학에 대한 인식 차이에 따라 다른 반응 등의 사례가 면담을 통해 발견되었다. 도구로 측정한 양적 자료와 학생들의 면담 내용인 질적 자료를 비교한 결과 스스로 반응한 양적 자료와 다르게 자신을 평가하는 학생들이 대부분이었다. 이는 다양한 특성을 지닌 학생들이 검사 도구와 상호작용하는 과정에서 비롯된다. 검사 도구와 관련된 불일치 유형은 '문항 개발자가 의도한 개념과 학생들이 이해하는 개념 간 차이'와 '표현된 반응과 속마음 간의 차이'로 나타났다. 검사 도구에서 사용한 용어가 학생들에게 모호하게 인식될 때 자의적이거나 일관성 없이 반응하는 경우가 전자에, 사회적 바람직성이나 자아 방어 기제에 의한 반응 왜곡은 후자에 해당한다. 이상 연구 결과를 바탕으로 자기보고식 검사 도구가 학생들의 실제 인식을 잘 반영하고 있는지 검토하고 정교화하려는 노력, 학습 경험을 획일적으로 고정시키는 평가 개선 등이 필요하다.

주제어 : 인지적 성취, 정의적 성취, 불일치 사례, 왜곡 반응, 자기보고식 검사 도구, 사회적 바람직성, 과학 교육 평가

References

Abd-El-Khalick, F., Summers, R., Said, Z., Wang, S., & Culbertson, M. (2015). Development and large-scale validation of an instrument to assess Arabic-speaking students' attitudes toward science. International Journal of Science Education, 37(16), 2637-2663.

Aikenhead, G. (2001). Students' ease in crossing cultural borders into school science. Science Education, 85, 180-188.

Aikenhead, G., & Jegede, O. J. (1999). Cross-cultural science education: A cognitive explanation of a cultural phenomenon. Journal of Research in Science Teaching, 36(3), 269-289.

Au, Y. (2007). A search on social desirability according to administered mode and demonstrable condition of a psychology testing. Journal of Educational Evaluation, 20(4), 235-258.

Bae, B., Lee, D., & Ham, K. (2015). Validation of the Korean short-version of social desirability scale(SDS-9) using the Rasch model. Korean Journal of Counseling, 16(6), 177-197.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. Thousand Oaks, CA: Sage.

Chi, E. (2011). Applying the Rasch model to explore the differences between countries for tests administered across countries. Journal of Educational Evaluation, 24(1), 89-106.

Cho, Y. (2003). A study on I-consciousness-we-consciousness-relationships between I-consciousness-we-consciousness and individuality relatedness, psychosocial maturity, and interpersonal problem. The Korean Journal of Counseling and Psychotherapy, 15(1), 91-109.

Cho, J., Kim, S., Kim, M., Ok, H., Lim, H., & Son, S. (2012). Ways of improving Korean students' affective characteristic based on PISA and TIMSS results. Seoul: Korea Institute for Curriculum and Evaluation.

Choe, S., Ku, J., Kim, J., Park, S., Oh, E., Kim, J., & Baek, H. (2013). Strategies for improving the affective characteristics of Korean students based on the results of PISA and TIMSS. Seoul: Korea Institute for Curriculum and Evaluation.

Choi, S. (2004). Social psychology of Korean people. The Korean Psychological Association, 2, 151-162.

Chung, S., & Shin, D. (2016). Trends of assessment research in science education. Journal of the Korean Association for Science Education, 36(4), 563-579.

Corbin, J., & Strauss, A. (2008). Basics of qualitative research: Techniques and procedures for developing grounded theory(3rd ed.). Thousand Oaks, CA: Sage.

Costa, V. (1995). When science is "another world": Relationships between worlds of family, friends, school, and science. Science Education, 79(3), 313-333.

Cronbach, L. (1946). Response sets and test validity. Educational and Psychological Measurement, 6(3), 475-494.

Crowne, D., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349-354.

Dudley, N., McFarland, L., Goodman, S., Hunt, S., & Sydell, E. (2005). Racial differences in socially desirable responding in selection contexts: Magnitude and consequences. Journal of Personality Assessment, 85(1), 50-64.

Ferrando, P., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. Educational and Psychological Measurement, 61, 997-1012.

Ferrari, J., Bristow, & Cowman, S. (2005). Looking good or being good? The role of social desirability tendencies in student perception of institutional mission and values. College Student Journal, 39(1), 7-13.

Fives, H., Huebner, W., Birnbaum, A., & Nicolich, M. (2014). Developing a measure of scientific literacy for middle school students. Science Education, 98(4), 549-580.

Ganster, D., Hennessey, H., & Luthans, F. (1983). Social desirability response effects: Three alternative models. Academy of Management Journal, 26(2), 321-331.

Heine, S., Lehman, D., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. Journal of Personality and Social Psychology, 82(6), 903-918.

Jűrges, H., Schneider, K., & Bűchel, F. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. Journal of the European Economic Association, 3(5), 1134-1155.

Keillor, B., Owens, D., & Pettijohn, C. (2001). A cross-cultural/cross-national study of influencing factors and socially desirable response biases. International Journal of Market Research, 43(1), 63-84.

Ku, J., Kim, S., Lee, H., Cho, S., & Park, H. (2016a). OECD Programme for International Student Assessment: An analysis of PISA 2015 results. Seoul: Korea Institute for Curriculum and Evaluation.

Ku, J., Kim, S., Lee, H., Cho, S., & Park, H. (2016b). OECD Programme for International Student Assessment: Establishing a foundation of PISA 2018 field trial. Seoul: Korea Institute for Curriculum and Evaluation.

Kim, S., Kim, K., & Park, J. (2014). The effect of mathematics achievement on changes in mathematics interest and values for middle school students. Journal of Research in Curriculum Instruction, 18(3), 683-701.

Kim, K., Kim, S., Kim, M., Kim, S., Kang, M., Park, H., & Jung, S. (2009). Comparative analysis of curriculum and achievement characteristics between Korea and high performing countries in PISA & TIMSS. Seoul: Korea Institute for Curriculum and Evaluation.

Kim, S. (2013). Measurement of social norms: An experimental study of response bias. Korean Political Studies, 22(2), 153-178.

Kim, S., Seo, H. (2011). Self-regulated learning ability related to science inquiry skill and affective domain of science in middle school students. Journal of Science Education, 35(2), 307-323.

Kim, M., & Cho, J. (2013). Analysis of the properties of affective achievement in science based on TIMSS and science teachers' perception. Journal of the Korean Association for Science Education, 33(1), 46-62.

Kim, S., Park, J., Kim, H., Jin, E., Lee, M., Kim, J., Ahn, Y., & Seo, J. (2012). Findings from TIMSS for Korea: TIMSS 2011 international results. Seoul: Korea Institute for Curriculum and Evaluation.

Kim, Y. (2010). Development of a social desirability scale(SDS-24). Journal of Korean Social Welfare Administration, 12(3), 1-39.

Kind, P., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. International Journal of Science Education, 29(7), 871-893.

Koballa, T. R. (1988). Attitude and related concepts in science education. Science Education, 72(2), 115-126.

Krynowsky, B. (1988). Problems in assessing student attitude in science education: A partial solution. Science Education, 72(4), 575-584.

Kwak, Y. (2017). Exploration of features of Korean eighth grade students' attitudes toward science. Journal of the Korean Association for Science Education, 37(1), 135-142.

Lee, J. (2016). Analysis of changes in the learning environments of middle school science classes. Journal of the Korean Association for Science Education, 36(5), 717-727.

Lee, M., & Kim, K. (2004). Relationship between attitudes toward science and science achievement. Journal of the Korean Association for Science Education, 24(2), 399-407.

Lee, M., Sohn, W., & No, U. (2007). The results from PISA 2006. Seoul: Korea Institute for Curriculum and Evaluation.

Lee, M., Park, S., Sohn, W., & Nam, M. (2007). Technical report for PISA 2006 main study. Seoul: Korea Institute for Curriculum and Evaluation.

Lee, M., Choi, J., Lee, J., & Shin, M. (2016). A preliminary study of defensive response style on a self-report personality assessment. The Journal of the Korean Association of Psychotherapy, 8(2), 61-80.

MaCrae, R., & Costa, P. (1983). Social desirability scales: More substance than style. Journal of Consulting and Clinical Psychology, 51, 882-888.

Martin, M., Mullis, I., & Foy, P. (2012). TIMSS 2011 international science report. MA: Boston College.

Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Willey(Eds.). Improving inquires in social science. Hillsdale, NJ: Erlbaum.

Ministry of Education Science, and Technology (2011). Science curriculum. Notification No. 2011-361 of MOEST. Seoul: MOEST.

Mullis, I., Martin, M., & Foy, P. (2012). TIMSS 2011 international mathematics report. MA: Boston College.

Murayama, K., Zhou, M., & Nesbit, J. (2009). A cross-cultural examination of the psychometric properties of response to the achievement goal questionnaire. Educational and Psychological Measurement, 69(2), 266-286.

Myeong, J., & Crawley, F. E. (1993). Predicting and understanding Korean high school students' science track choice: Testing the theory of reasoned action by structural equation modeling. Journal of Research in Science Teaching, 30(4), 381-400.

Ones, D., Viswesvara, C., & Reiss, A. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. International Journal of Science Education, 25(9), 1049-1079.

Osborne, J., Simon, S., & Tytler, R. (2009). Attitudes towards science: An update. Paper presented at the annual meeting of the American Educational Research Association, San Diego, California.

Park, C. (2007). The trend in the Korean middle school student's affective variables toward mathematics and its effect on their mathematics achievement. The Mathematical Education, 46(1), 19-31.

Park, C. Jeong, E., Kim, K., Han, K., Jun, H., & Lee, S. (2004). Teachers, instruction, and achievement based on TIMSS 1999. Seoul: Korea Institute for Curriculum and Evaluation.

Park, H., (2008). Test of group invariance for the structural model among motivation, self-concept and student achievement: Using PISA 2006 data. Journal of Educational Evaluation, 21(3), 43-67.

Paulhus, D. (1984). Two component models of socially desirable responding. Journal of Personality and Social Psychology, 46, 589-609.

Paulhus, D. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? Journal of Personality and Social Psychology, 74, 1197-1208.

Reise, S., & Flannery, W. (1996). Assessing person-fit on measures of typical performance. Applied Measurement in Education, 9(1), 9-26.

Saldaña, J. (2009). The coding manual for qualitative researchers. Thousand Oaks, London: Sage.

Sang, K., Kwak, Y., Park, J., & Park, S. (2016). The Trends in International Mathematics and Science Study (TIMSS): Findings from TIMSS 2015 for Korea. Seoul: Korea Institute for Curriculum and Evaluation.

Schluf, Boaz, Hattie, J., & Dixon, R. (2008). Factors affecting responses to Likert type questionnaires: Introduction of the ImpExp, a new comprehensive model. Social Psychology of Education, 11(1), 59-78.

Schunk, D., & Pajares, F. (2009). Self efficacy theory. In Wentzel, K., & Wigfield, A. (Eds). Handbook of motivation at school. New York: Routledge.

Seo, J., Choi, J., & Kim, Y. (2007). Comparison of life learning skills of gifted science students and normal students in high school. Biology Education, 35(1), 61-72.

Shen, C., & Pedulla, J. (2000). The relationship between students' achievement and their self-perception of competence and rigour of mathematics and science: A cross-national analysis. Assessment in Education, 7(2), 237-253.

Shin, H., & Sohn, W. (2014). Applying a mixed Rasch model to investigate response styles in TIMSS 2011 math enjoyment scale. Journal of Educational Evaluation, 27(2), 429-448.

Sohn, W. (2017). Individual difference and consistency in response scale use. Korean Journal of Educational Research, 55(1), 23-43.

Son, E., Cha, J., & Kim, A. (2007). Test of construct equivalence of personality inventory in low and high socially desirable responding groups. Korean Journal of Social and Personality Psychology, 21(2), 71-87.

Song, H. (2010). Development of a self-reported executive function rating scale for the Korean high school students: A preliminary study. The Korean Journal of Clinical Psychology, 29(1), 109-124.

Stephen, A. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. Educational and Psychological Measurement, 60(3), 340-360.

Stöber, J. (2001). The social desirability scale-17(SDS-17): Convergent validity, discriminant validity, and relationship with age. European Journal of Psychological Assessment, 17(3), 222-232.

Strauss, A. L. (1987). Qualitative analysis for social scientists. Cambridge: Cambridge University Press.

Suzuki, L., & Ponterotto, J. (2007). Handbook of multicultural assessment: Clinical, psychological, and educational applications. San Francisco, CA: John Wiley & Sons, Inc.

Vailant, G. (1992). Ego mechanism of defense. Washington: American Psychiatric Press.