중학생 과학탐구활동 수행평가 시 총체적 채점에서 나타나는 채점자간 불일치 유형 분석

김형준 · 유준희*
기산중학교 · 서울대학교

An Analysis on Rater Error in Holistic Scoring for Performance Assessments of Middle School Students' Science Investigation Activities

Kim, Hyung Jun · Yoo, Junehee*
Gisan Middle School · Seoul National University

Abstract: The purpose of this study is to understand raters' errors in rating performance assessments of science inquiry. For this, 60 middle school students performed scientific inquiry about sound propagation and 4 trained raters rated their activity sheets. Variance components estimation for the result of the generalizability analysis for the person, task, rater design, the variance components for rater, rater by person and rater by task are about 25%. Among 4 raters, 2 raters' severity is higher than the other two raters and their severities were stabilized. Four raters' rating agreed with each other in 51 cases among the 240 cases. Through the raters' conferences, the rater error types for 189 disagreed cases were identified as one of three types; different salience, severity, and overlooking. The error type 1, different salience, showed 38% of the disagreed cases. Salient task and salient assessment components are different among the raters. The error type 2, severity, showed 25% and the error type 3, overlooking showed 31%. The error type 2 seemed to have happened when the students responses were on the borders of two levels. Error type 3 seemed to have happened when raters overlooked some important part of students' responses because she or he immersed her or himself in one's own salience. To reduce the above rater errors, raters' conference in salience of task and assesment components are needed before performing the holistic scoring of complex tasks. Also raters need to recognize her/his severity and efforts to keep one's own severity. Multiple raters are needed to prevent the errors from being overlooked. The further studies in raters' tendencies and sources of different interpretations on the rubric are suggested.

Key words: rater error, salient assessment component, severity, overlook, performance assessment, science investigation activity, middle school student

Ⅰ. 서 론

과학탐구활동에 대한 분석적 평가와 총체적 평가 사이의 갈등은 오래된 문제 중의 하나이다. 총체적 채점은 수행과제의 전체적인 인상에 의해 단일 점수를 부여하는 방법이고 분석적 채점은 몇 가지 하위요소로 구성된 평가틀에 따라 채점하여 하위요소의 점수를 총합하는 방법이다(Klein *et al.*, 1998; 김명숙, 1999; 지은림, 1999). 총체적 채점은 수행과제의 부분적인 고려보다는 전체적인 과제의 특성에 더 큰 의미를 부여하며(이규민, 2007) 과제 전체의 의미가 부분적인 평가요소들의 합보다 더 큰 과제인 경우에 사용된다(Klein *et al.*, 1998). 따라서 총체적 채점을 할 때는 수준별로 학생들의 특징을 가장 잘 반영하는 채점 기준을 설정하는 것이 중요하나(Waltman & Koency, 1998), 여전히 총체적 채점 기준의 모호성에 대한 논의에서 벗어나기는 어렵다. 과학탐구활동에서 달성되어야 할 목표를 명확하게 서술한 분석적인 평가틀(APU, GCSE등)이 개발되면서 평가요소에 대한 불명확성은 사라졌지만, 전체가 부분의 합 이상의 의미가 없다는 가정 속에 작은 부분들의 합으로 과학 활동을 축소하여 탐구 과정의 전체적인 의미를 살

*교신저자: 유준희(yoo@snu.ac.kr)
**2011.10.26(접수) 2011.12.08(1심통과) 2012.01.13(2심통과) 2012.01.16(최종통과)

릴 수 없다는 비판과 평가틀이 분석적이어서 총체적인 탐구활동을 제한한다는 문제가 제기되어왔다(Woolnough, 1989). 분석적인 탐구기능의 평가틀은 실제의 복합적 상황에서 이루어지는 총체적인 탐구활동을 평가하는데 타당도를 담보할 수 없으며, 제한적인 가치를 가진다는 것이다(Black, 1990).

수행평가 시 채점방식에 따른 신뢰도 분석에 대한 연구보고에 따르면(김형준, 2010), 총체적 채점 방식의 채점자 내적 일치도는 0.77~0.81로 분석적 채점의 0.72~0.77보다 높게 나타났다. 분석적 채점 방식에서 채점척도를 3등급으로 한 경우, 채점자간 일치도는 0.85~0.92로 총체적 채점 방식의 0.71~0.90보다 높게 나타났다. 따라서 복합적인 탐구활동을 보다 타당하게 평가하기 위해 총체적 채점방식의 도입하는 경우에는 채점의 신뢰도를 확보가 필요하며, 이를 위해서는 채점자간 일치도를 높이기 위한 기초 연구가 필요하다. 이에 본 연구는 총체적 채점시 나타나는 채점자간 불일치의 정도와 유형을 질적으로 분석하여 채점자간 불일치를 어떻게 효과적으로 제어할 것인지에 대한 시사점을 제공하고자 한다.

채점 시 오차는 크게 채점자, 과제, 학생 및 각 분산성분 사이의 상호작용에 의해 발생한다(Clauser *et al.*, 1999; Clauser *et al.*, 2006). 일반화 가능도 이론에서 학생을 제외한 분산성분은 오차를 나타낸다고 가정한다(Clauser *et al.*, 2006). 분산성분으로는 학생(p), 채점자(o), 과제(i), 학생과 채점자(po), 학생과 과제(pi), 채점자와 과제(oi) 및 혼합분산(poi)가 존재한다. 분산 성분의 하나로서 채점자와 관련된 오차(rater related error)에 대한 연구가 있어 왔으며(Causer *et al.*, 1999; Clauser *et al.*, 2006), 수행평가 채점의 신뢰도에 영향을 주는 요인으로 채점자 특성, 채점자간 일관성 및 평정자간 엄격성 차이 등에 대한 논의가 지속적으로 있어왔다(설현수, 2010; 송미영 외, 2009; 지은림, 2008). 즉, 어떤 채점자는 다른 채점자보다 더 관대할 수도 있으며, 채점자마다 어떤 응답이 다른 응답보다 더 좋다고 다르게 판단할 수 있다. 이러한 채점자간 불일치는 채점신뢰도를 저하시키는 경향이 있다(Klein *et al.*, 1998). 채점자간 불일치는 채점자들과 연관된 다양한 요인들에 의해서 발생하는데, 채점자의 피로, 채점표 제작자 및 출제자의 의도에 대한 이해 부족, 해독하기 어려운 학생들의 글씨체 등이 주는 왜곡, 학생의 선행 응답의 질에 의한 왜곡 등이 이유가 제시되고 있다(Wilson & Case, 1997).

채점 불일치에 대한 선행연구를 분석한 결과, 채점자마다 각기 다른 등급으로 학생들의 답안을 평가하는 원인은 다음과 같이 여섯 가지로 분류할 수 있다(Guilford, 1954; Myford & Wolfe, 2003; Polio, 1997; Wilson & Case, 1997). 첫 번째 유형은 채점자간 중요하게 생각하는 부분의 차이 때문에 발생하는 후광효과이다(Myford & Wolfe, 2003). 후광효과는 많은 평가요소들이 하나의 중요한 평가요소에 의해 영향을 받아 결정되는 것을 의미한다(Guilford, 1954). 즉 채점자가 중요하게 생각하는 평가요소에 대한 평가가 채점자가 덜 중요하게 생각하는 평가요소에 대한 평가에 직접적으로 영향을 미치는 경우가 대표적인 예이다(Myford & Wolfe, 2003). 두 번째 유형은 채점자의 엄격함과 관대함에 의해 발생하는 불일치로 채점자가 채점 시 학생들이 일관되게 받아야 할 점수보다 더 박하게 주거나 후하게 주는 것을 의미한다(Wilson & Case, 1997). 관대함은 학생들이 받아야 할 점수보다 더 많이 채점자가 점수를 부여하는 경향이다. 이것은 채점 결과를 해석하는데 문제를 야기한다. 학생들이 높은 점수를 받을 때 우리는 높은 점수를 주는 채점자의 경향 때문인지 학생들의 수행이 잘되어 높은 점수를 받았는지 알 수가 없다. 반대로 낮은 점수를 주는 채점자도 있는데 이를 엄격한 채점자라고 부른다. 엄격한 채점자 역시 관대한 채점자만큼 결과에 유효한 영향력을 만드는 문제를 일으킬 수 있다. 엄격하고 관대한 채점자 모두 유사한 종류의 문제를 일으키기 때문에 연구자들은 채점자들이 너무 높거나 낮게 채점하는 경향을 관대함의 불일치라고 부른다(Myford & Wolfe, 2003). 관대함의 차이에서 오는 불일치는 최근 다국면 라쉬 측정모형 등을 이용하여 점수를 보정하는 방안이 측정평가연구에 모색되고 있다(설현수, 2010). 세 번째 유형은 채점자의 실수로 인해 발생하는 불일치이다. 채점자의 실수는 채점 결과의 재검 등을 통해서 보완될 수 있는 부분이기 때문에 이와 관련된 연구 결과가 보고된 적은 거의 없다. 그러나 현실적으로 발생하고, 어떤 실수가 발생하는지를 파악하기 위하여 본 연구에서는 하나의 유형으로 분류하였다. 네 번째 유형은 학생의 글씨체가 보기 어렵게 작성 되었을 때 채점자들이 알아볼 수 있는 정도가 각기 달라 발생하는 불일치이다(Polio, 1997).

다섯 번째 유형은 불성실하게 답변한 학생의 답안이 있을 때 채점자들 중에는 더 낮은 수준으로 채점하는 경향 때문에 발생하는 불일치이다(Polio, 1997). 여섯 번째 유형은 학생 답안의 의미가 불명확할 때는 채점자들의 주관적인 해석에 따라 다른 채점 척도를 부여하기 때문에 발생하는 불일치이다(Polio, 1997).

채점자의 특성이 평가의 신뢰도에 미치는 영향은 주로 측정평가 영역에서 양적 연구로 이루어지고 있다(Clauser *et al.*, 1998; Clauser *et al.*, 1999; Clauser *et al.*, 2006; 설현수, 2010; 송미영 외, 2009; 지은림, 2008). 이와 같은 연구들은 통계적 보정을 통해 채점자간 불일치를 제어하는데 도움을 줄 수 있지만, 불일치가 발생하는 원천적인 원인을 파악하여 불일치를 제어하는데는 도움을 주기 어렵다. 이에 본 연구에서는 중학생의 과학 수행평가 시 총체적 채점에서 나타나는 채점자간 불일치의 정도와 유형을 보다 질적으로 분석하여 채점자간 불일치에 대하여 보다 원천적으로 제어할 수 있는 방안을 제시하고자 한다. 본 연구의 구체적인 연구 질문은 다음과 같다.

첫째, 중학생의 과학탐구활동 수행평가 시 총체적 채점에서 채점자간 불일치의 정도와 유형은 어떻게 나타나는가?

둘째, 중학생의 과학탐구활동 수행평가 시 채점자가 중요하게 생각하는 요소가 다르기 때문에 발생하는 불일치는 어떤 양상을 보이는가?

셋째, 중학생의 과학탐구활동 수행평가 시 채점자의 관대함과 엄격함의 경향에 다르기 때문에 발생하는 불일치는 어떤 양상을 보이는가?

넷째, 중학생의 과학탐구활동 수행평가 시 채점자의 실수에 의해 발생하는 불일치는 어떤 양상을 보이는가?

Ⅱ. 연구 방법 및 절차

1. 연구 참가자

본 연구의 참가자는 모두 4명으로 물리 전공의 현직 과학 교사로 이들의 배경 정보는 Table 1과 같다. 연구 참가자들의 중등학교 교직경력은 3년 이상이며, 모두 학교 현장에서 수행평가 및 채점의 경험이 있다. 채점자 1은 본 연구진 중의 한사람이다.

Table 1
연구 참가자

	전공	경력	성별
채점자1	물리	9년	남
채점자2	물리	3년	남
채점자3	물리	4년	남
채점자4	물리	4년	남

2. 연구 과정 및 방법

가. 채점 기준의 작성 및 채점

연구진은 중학생을 위한 과학탐구 수행평가 과제를 개발하였으며, 총체적 채점 기준안를 구성하였다. 수행평가가 실시된 후 채점자1이 가채점하여 기작성한 채점 기준안을 수정하였다. 본 연구에서는 작성한 채점 기준안의 타당도를 검토하기 위하여 채점자1과 2는 채점 기준안과 가채점한 학생 응답을 대조하면서 의견을 조율하고 필요한 경우 채점 기준안을 수정하였다. 이후 채점자1과 2는 수정된 채점기준과 각 등급에 해당하는 학생응답의 예시를 공유하고 채점 기준에 숙달되도록 4시간 정도 연습하였다. 채점자3과 4는 채점자2와는 별도로 협의회를 가졌으며, 채점자1과 2의 협의에 의해 재수정된 채점 기준과 각 등급에 해당하는 학생응답 예시에 대하여 채점자1과 협의 및 연습을 수행한 후 본 채점을 실시하였다.

나. 채점자간 불일치 유형의 분류

연구진은 선행 연구 고찰을 통해 Table 2와 같이 채점자간 불일치 유형 분류틀을 작성하였다(Guilford, 1954; Myford & Wolfe, 2003; Polio, 1997; Wilson & Case, 1997). 채점 결과에 대한 통계적 분석을 완료한 후, 4명의 채점자가 모여서 채점자간 불일치가 발생한 학생 답안을 검토하면서 선행연구를 바탕으로 작성한 채점자간 불일치 유형 분석틀에 근거하여 채점이 불일치한 원인을 판단하여 채점자간 불일치 유형을 분류하였다. 이 때, 채점자간 불일치 유형을 분류하면서 채점자간의 논의를 통해서 본 연구 대상에 적절하게 분류틀을 Table 3과 같이 수정하였다. 예를 들어 빈도수가 적게 나타나는 학생의 읽기 어려운 글씨체에 의해 발생하는 불일치, 학생의 불성실한 응답에 의해 발생하는 불일치, 학생들의

모호한 서술에 의해 발생하는 불일치 등 학생에게 귀인되는 원인은 단순 채점 오류로 단일화하였으며, 이에 대한 분석은 하지 않았다. 세 번째 불일치 유형인 채점자의 실수는 본 연구에서는 채점자가 중요하게 생각하지 않은 채점요소를 간과해서 발생하는 불일치로 특징화할 수 있었다. 즉, 한 채점자가 실수를 하였을 때, 그 이유를 논의한 결과 채점자가 중요하게 생각하는 부분에 집중하여 채점을 하다보면 다른 부분은 간과한 채 학생 답안의 질을 결정하기 때문인 것으로 논의되었다.

Table 2
협의전 채점자 관련 불일치 유형의 분류틀

분류번호	불일치 유형
1	채점자간 중요하게 생각하는 부분의 차이 때문에 발생하는 불일치(두드러진 요소의 모형)
2	경계에서의 모호함때문에 발생하는 불일치 (엄격함과 관대함에 관한 채점자의 성격)
3	한 채점자의 실수에 의한 불일치
4	읽기 어려운 학생의 글씨체에 의해 발생하는 불일치
5	학생의 불성실한 응답에 의해 발생하는 불일치
6	학생들의 모호한 서술에 의해 발생하는 불일치

Table 3
협의후 채점자 관련 불일치 유형의 분류틀

분류번호	불일치 유형
1	채점자간 중요하게 생각하는 부분의 차이 때문에 발생하는 불일치
2	엄격함과 관대함에 의해 발생하는 불일치
3	채점자가 중요하게 생각하지 않은 부분에서 간과하기 때문에 발생하는 불일치
4	단순한 채점 불일치

채점자간 불일치가 발생한 사례에 대하여 채점자간 불일치 유형을 파악하기 위하여 2차례에 걸쳐서 협의회를 가졌다. 첫 번째 협의회는 4명의 채점자가 모여서 과제1과 과제2에 대한 응답 중 채점자간 불일치가 나타나는 응답에 대하여 3시간 동안 진행하였다. 협의회 동안 채점자들은 각각의 채점자간 불일치 응답 사례에 대하여 자신의 채점 근거를 설명하였고, 그러한 설명을 바탕으로 채점자끼리 서로 다른 채점을 하게 된 원인을 파악하여 채점자간 불일치 유형에 대한 의견을 수렴하였다. 이러한 과정에서 채점자간 불일치 유형 분류틀을 다시 수정하였고, 각각의 사례에 대한 채점자간 불일치 유형에 대하여 최종적인 합의를 할 수 있었다. 일주일 뒤에 두 번째 협의회를 진행하여 과제3과 과제4에 대한 학생 응답 중 채점자간 불일치가 나타나는 경우에 대하여 채점자간 불일치 유형을 판단하여 분류하였다.

3. 분석도구

문항반응이론을 이용하여 문항의 양호도를 확인하였고, 평가의 신뢰도 분석을 위해서는 채점자 내 신뢰도와 채점자 간 신뢰도를 검토하였다(김형준 외, 2010; 성태제, 2005). 기초통계 분석프로그램으로 spss 12.0을 사용하였고, 문항반응이론에 기반한 분석프로그램으로 미국 버클리 대학 평가 및 측정 센터(Berkely Evaluation and Assessment Research Center, BEAR)에서 개발한 Grade Map을 사용하였다(Wilson & Sloane, 2000). Grade Map을 이용하여 일반화가능도 이론에 의한 분산성분별 분산 추정치를 구하여 채점자에 의한 분산의 크기를 추정하였으며, 과제간 채점 결과를 표준화하여 채점자의 관대함과 엄격함의 경향을 파악할 수 있다.

4. 과학탐구활동 수행평가 도구와 평가기준

본 연구를 위하여 서울 소재 영재원에서 중학생 1학년 20명, 2학년 22명, 3학년 18명 등 총 60명을 대상으로 소리의 전달과정을 공기입자모형으로 설명하는 탐구활동을 수행평가로 진행하였다. 탐구활동은 과제1, 과제2, 과제3, 과제4로 이루어져 있고 각각의 과제들은 P-O-E(예상-관찰-설명)의 단계로 구성되었다. 과제1은 빨대피리를 불어서 큰 소리와 작은 소리, 높은 소리와 낮은 소리가 전달될 때 공기의 움직임을 예상하고 녹음한 소리의 파형을 관찰한 후 설명하는 것이다. 과제2는 파동 용수철을 사용하여 진폭이 큰 신호와 작은 신호, 1초 동안 신호를 많거나 적게 만들어 보내는 활동에서 용수철의 각 부분의 움직임을 예상, 관찰, 설명하게 하였다. 과제3은 시뮬레이션을 작동시켜서 큰 소리와 작은 소리, 높은 소리와 낮은 소리에 대한 공기 입자의 움직임을 예상, 관찰, 설명하게 하였다. 마지막으로 과제4는 excel을 활용한 프로

그램으로 직접 큰 소리와 작은 소리, 높은 소리와 낮은 소리를 만들어 듣고 최종적으로 공기입자의 움직임으로 소리의 특성을 설명하게 하였다. 학생들은 조별로 과제를 수행하고 개인별로 활동지를 작성하였으며, 학생들이 작성한 활동지를 수합하여 채점하였다.

평가목표는 과학 현상을 관찰과 측정된 증거들을 사용하여 설명하는 능력이며 총체적 방식으로 채점하였다. 총체적 채점 방식의 평가 수준에 관련한 선행연구 중 Etkina *et al.* (2006)은 과학 수행평가의 수준을 공란, 부적절함, 향상이 필요함, 적절함으로 4수준으로 나누었고, Halonen *et al.*(2003)은 과학탐구 평가에 5수준의 채점척도를, Hafner *et al.*(2003)는 과학 탐구과제의 구두발표 평가 시 5수준의 채점척도를 사용하였다. 본 연구에서는 선행 연구 결과 및 학교에서의 활용을 고려하여 총체적 채점 방식의 채점척도를 전문가적, 진보, 보통, 초보, 설명 못함 등 5수준으로 설정하였다. 본 연구에서 사용한 과제 목표에 다른 평가 목표 및 채점기준표를 Table 4에 제시하였다.

Ⅲ. 연구 결과 및 논의

1. 과학 수행평가에 대한 총체적 채점 결과의 불일치 정도

가. 일반화 가능도 이론에 의한 분산 성분별 분산 추정치

일반화 가능도 이론에 의한 분산분석 결과를 Table 5에 제시하였다. 분산분석결과에 의하면, 채점 결과 중 46.5%는 학생의 능력으로, 21.9%는 학생과 채점자의 상호작용으로, 24.9%는 학생, 채점자, 과제의 상호작용으로 설명할 수 있다. 채점자와 관련된 분산성분인 채점자, 학생과 채점자의 상호작용, 채점자와 과제의 상호작용으로 각각 전체 분산의 2.6%, 21.9%, 1.9% 등 총 25.4%를 추정할 수 있다. 과제와 관련된 분산성분인 과제, 학생과 과제의 상호작용, 채점자와 과제의 상호작용으로 각각 전체 분산의 1.2%, 1.0%, 1.9% 등 총 4.1%를 추정할 수 있었다. 즉, 채점자 관련 분산성분으로 추정할 수 있는 분산의 크기가 과제 관련 분산성분으로 추정할 수 있는 분산의 크기보다 크다.

Table 4
과제 목표 및 채점 기준표

과제 목표				수 준	평가 기준
과제 1	과제 2	과제 3	과제 4		
빨대 피리에서 발생한 소리의 전달과정을 소리의 특성에 따라 파형을 분석하여 설명할 수 있는 능력	보내는 신호에 따라 용수철의 움직임을 설명할 수 있는 능력	시뮬레이션을 통하여 소리가 전달되는 동안 공기입자의 움직임을 설명 수 있는 능력	다양한 진동수와 진폭의 소리 파형을 예상하고, 시뮬레이션을 이용하여 관찰하며, 이를 종합하여 소리가 전달될 때 공기입자 전체의 움직임과 입자 하나의 움직임을 설명할 수 있는 능력	전문가적	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었으며 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되었다.
				진보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.
				보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.
				초보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.
				설명 못함	예상하기, 측정하기, 설명하기의 각 부분의 설명이 대부분 과학적으로 타당하지 않게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.

분산성분 중 채점자와 과제의 상호작용(oi)은 채점자의 엄격함과 관대함에 의해 발생하는 변동성을 대표한다고 알려졌다(Harick, *et al.*, 2009). 나머지 분산성분과 채점자간 불일치 원인유형에 대해서 아직 알려진 바가 없다.

나. 채점자별 관대함 및 엄격함의 경향

문항반응이론을 기반으로 하는 Grade Map을 통해 각 채점자별 점수를 표준화하였고 그 결과를 Figure 1에 제시하였다. 채점자 1, 2는 채점자 4명의 평균 점수보다 낮은 점수를 주는 엄격한 채점자로, 채점자 3, 4는 평균점수보다 높은 점수를 주는 관대한 채점자로 해석할 수 있다.

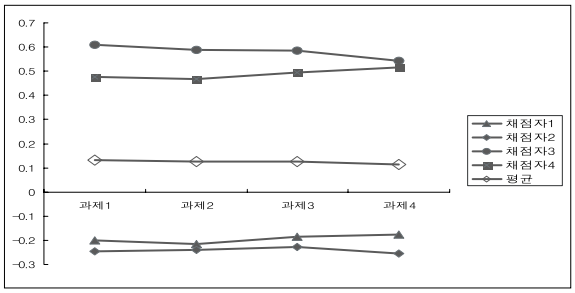


Fig. 1 *채점자별 과제의 평균 점수*

다. 채점자간 일치 빈도수와 불일치의 정도

Table 6은 과제별로 채점결과가 일치한 사례의 빈도수를 나타내는데 전체 사례 240건 중 채점자 4명이 모두 동일한 점수를 준 사례는 51건이었고 나머지 199건이 한명이라도 다른 점수를 준 사례였다. 이 중에서 3인 일치는 80건으로 제일 많았고, 2인 2쌍 일치가 58건, 2인 1쌍 일치가 49건, 일치 없음이 2건이었다. 과제별로 보면 과제1, 2, 3은 3인 일치가 가장 많았고 과제4는 2인 2쌍일치가 가장 많았다. 이는 과제4가 앞의 세 가지 과제보다 좀 더 종합적인 과제의 성격을 가졌기 때문에 채점자마다 중요시하는 요소가 다르게 나타났기 때문으로 해석된다. 이에 대한 세부적인 분석은 다음 절에서 제시한다.

채점자 일치 정도에 따른 점수 차이를 Table 7에 제시하였다. 채점 결과가 3인이 일치하는 경우의 87.5%와 2인 2쌍으로 일치하는 경우의 96.6%가 1수준의 차이를 나타냈으며, 2인 1쌍으로 일치하는 경우는 채점 결과가 2수준 차이가 나는 비율이 87.8%, 4명의 채점자가 모두 다르게 채점한 경우인 일치 없음에서는 3수준 이상 차이가 나는 비율이 100%이다. 이는 채점자 간의 일치되는 정도가 작아질수록 채점 결과의 수준 차이가 커지는 것으로 해석할 수 있다.

라. 채점자가 인식한 불일치 유형의 분포

협의회를 통해 4명의 채점자가 합의한 채점자간 불일치 유형의 빈도수를 과제별로 Table 8에 제시하였다. 총 240개의 사례 중 채점자간 불일치가 발생한 사례는 189사례이고, 이중 72사례(38.1%)가 채점자가 중요하게 생각하는 부분이 다르기 때문에 발생하는 유형1에 의해 불일치가 발생하였다고 채점자들이 인식하였다. 불일치 유형 중 이 38.1%로 가장 높게 나타났다. 또한 엄격함과 관대함의 차이에 의한 유형2에 의해 발생한 불일치 사례는 47사례(24.9%), 중요하게 생각하는 부분에 집중하다보니 그 외의 부분을 간과하여 발생하는 유형3에 의한 불일치는 58사례(30.7%)로 나타났다.

Table 5
일반화 가능도 이론 $p \times I \times o$ 설계에 대한 분산 분석표

분산성분	자유도	제곱합	평균제곱	분산추정치(%)
학생(p)	59	708.80	12.01	0.65 (46.5)
채점자(o)	3	32.44	10.81	0.04 (2.6)
과제(i)	3	21.95	7.32	0.02 (1.2)
학생과 채점자(po)	177	277.86	1.57	0.31 (21.9)
학생과 과제(pi)	177	70.87	0.40	0.01 (1.0)
채점자와 과제(oi)	9	17.38	1.93	0.03 (1.9)
혼합분산(poi)	531	184.55	0.35	0.35 (24.9)
전체	959	1,313.85		1.41(100.0)

Table 6
과제별 채점자 일치 정도 빈도수(백분율)

채점자 일치 정도	과제1	과제2	과제3	과제4	계
4인 일치	13(21.7)	14(23.3)	13(21.7)	11(18.3)	51(21.3)
3인 일치	23(38.3)	21(35.0)	21(35.0)	15(25.0)	80(33.3)
2인 2쌍 일치	12(20.0)	14(23.3)	12(20.0)	20(33.3)	58(24.2)
2인 1쌍 일치	11(18.3)	11(18.3)	14(23.3)	13(21.7)	49(20.4)
일치 없음	1(1.7)	0(0)	0(0)	1(1.7)	2(0.8)
계	60(100)	60(100)	60(25.00)	60(100)	240(100)

Table 7
채점자 일치 정도에 따른 점수의 차이별 빈도수(백분율)

채점자 일치 정도 / 채점결과의 차이	3인 일치	2인 2쌍 일치	2인 1쌍 일치	일치 없음	계
1수준 차이	70(87.5)	56(96.6)	0(0)	0(0)	126(66.7)
2수준 차이	10(12.5)	2(3.4)	43(87.8)	0(0)	55(29.1)
3수준 이상 차이	0(0)	0(0)	6(12.2)	2(100)	8(4.2)
계	80(100)	58(100)	49(100)	2(100)	189(100)

Table 8
과제별 채점자가 인식한 채점자간 불일치 유형 빈도수(백분율)

불일치 유형	과제1	과제2	과제3	과제4	계
1. 중요시하는 부분의 차이	19(40.4)	25(54.3)	13(27.7)	15(30.6)	72(38.1)
2. 관대함의 차이	13(27.7)	5(10.9)	12(25.5)	17(34.7)	47(24.9)
3. 간과하는 내용	12(25.5)	14(30.4)	17(36.2)	17(34.7)	58(30.7)
4. 단순 채점자 오차	3(6.4)	2(4.3)	5(10.6)	2(4.3)	12(4.3)
계	47(100)	46(100)	47(100)	49(100)	189(100)

2. 불일치 유형1: 채점자간 중요시하는 요소가 다르기 때문에 발생하는 불일치

채점자 간에 중요하게 생각하는 평가 대상이 다르기 때문에 발생하는 불일치 유형1에 의해 채점 결과가 얼마나 달라지는지를 Table 9에 제시하였다. 유형1에 의하여 채점 결과가 1수준의 차이가 나는 경우가 총 72사례 중 47사례(65%)로 가장 많이 나타났으며, 보통과 초보 사이가 18회, 진보와 보통 사이가 14회로 나타났다. 과제별로 보면, 주로 과제 1과 과제 2의 채점에서 많이 발생하였다.

불일치 유형1은 채점자가 중요시하는 대상에 따라 두 가지 유형으로 나눌 수 있었다. 첫 번째는 채점자마다 중요하게 생각하는 과제 요소가 다른 경우이고, 두 번째는 채점자마다 중요하게 생각하는 평가 요소가 다른 경우이다. 각각의 사례를 다음에 제시하였다.

가. 채점자마다 중요하게 생각하는 과제 요소가 다른 경우

채점자간 불일치 유형1에 의하여 채점자간 불일치가 발생하는 경우 중 과제에서 중요하게 생각하는 요소가 다른 경우의 사례는 다음과 같다.

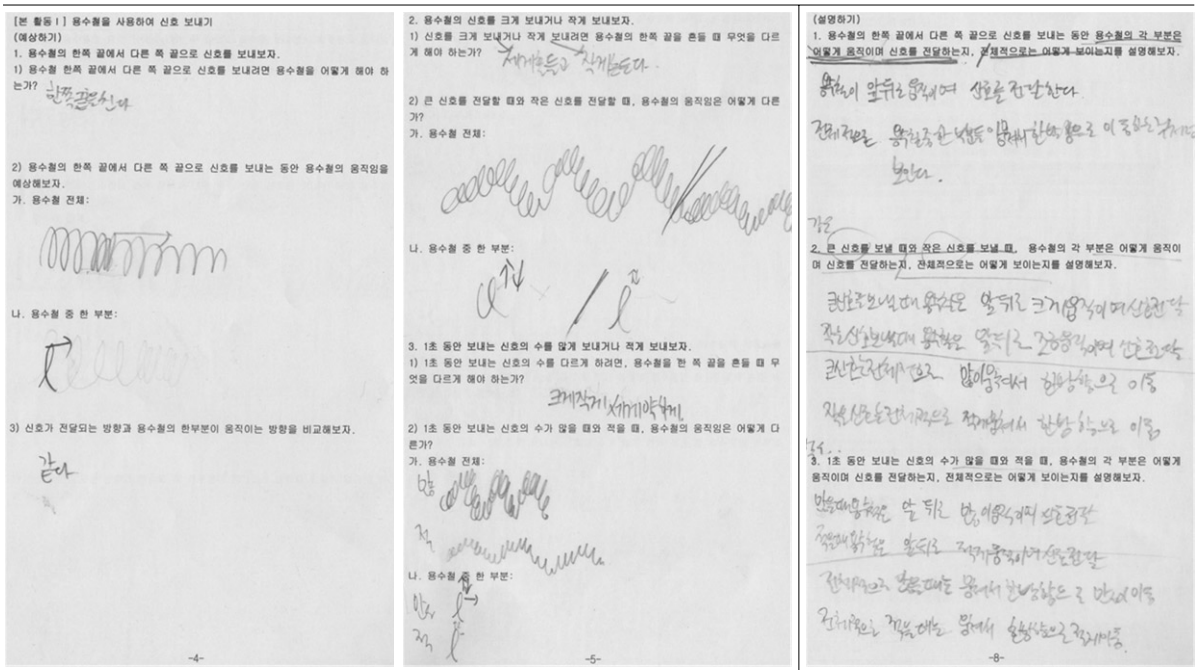
사례 1 : 과제2 요소 중 예상하기에서 일관성을 중요시 하는 채점자 1

Figure 2의 학생 응답에 대하여 채점자 1, 2, 3, 4

Table 9
불일치 유형1에 의한 채점 결과의 차이

채점 결과의 차이		과제1	과제2	과제3	과제4	계	
1수준 차이	전문가적 ↔ 진보	2	1	–	3	6	47
	진보 ↔ 보통	5	4	5	–	14	
	보통 ↔ 초보	5	8	3	2	18	
	초보 ↔ 설명못함	2	6	–	1	9	
2수준 차이	전문가적 ↔ 보통	1	3	1	2	7	23
	진보 ↔ 초보	3	1	3	3	10	
	보통 ↔ 설명못함	1	2	–	3	6	
3수준 이상차이	전문가적 ↔ 초보	–	–	1	1	2	2
	진보 ↔ 설명못함	–	–	–	–	0	
	전문가적 ↔ 설명못함	–	–	–	–	0	
계		19	25	13	15	72	

는 각각 초보, 보통, 보통, 보통의 수준으로 판단하였다. 즉 채점자 1이 다른 3명의 채점자와 다르게 채점한 경우이다. 각 채점자들은 자신의 판단에 대하여 Table 10과 같이 설명하였다. 채점자 1의 경우는 관찰하기나 설명하기에서 일정하게 종파로 응답한 것보다는 예상하기에서 일정하게 응답하지 않는 것을 중요시하게 받아들여서 해당 응답을 초보로 판단하였다. 채점자 2, 3, 4의 경우는 관찰하기나 설명하기에서 일정하게 종파로 응답한 것을 중요하게 받아들여서 보통으로 판단하게 되는 근거가 되었다. 채점자 1



과제2의 예상하기 응답 | 과제2의 설명하기 응답

Fig. 2 *예상하기에서 일관성이 부족하나 설명하기에서 일관성이 나타난 응답 예시*

Table 10
Figure 2.의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
초보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자1 : 예상하기는 맞고 틀리는 것이 중요하지 않지만 신호를 보내는 동안의 움직임을 예상해서 설명하는 모형을 동일한 것으로 해야 한다. 학생 답안의 경우 일부는 종파, 일부는 횡파로 설명했기 때문에 초보의 수준으로 보아야 한다.
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자2 : 하지만 이후에 관찰하기나 설명하기에서는 일정하게 종파로 설명하였다. 그래서 보통의 수준으로 주었다. 채점자3 : 설명하기의 학생 답안을 보면 일부 서술들이 부정확하게 서술되어서 보통의 수준정도가 적당하다고 생각했다. 채점자4 : 예상하기 부분은 채점에 중요한 부분을 차지하고 있지 않기 때문에 보지 않고 관찰하기와 설명하기를 보면 학생의 답안은 보통의 수준정도 된다고 생각한다.

이 "예상해서 설명하는 경우 동일한 모형으로 해야 한다"라고 설명한 것으로 보아 제시된 평가기준을 보고 예상하기에서 일관성있게 설명하지 않은 부분을 부각하여 받아들인 것으로 해석된다. 또한 제시된 평가 기준에서 보통 수준은 '...충분하게 서술...' 이고, 초보 수준은 '매우 부족...' 이지만, 채점자 1은 본인의 채점에 대해서 설명할 때는 "일관성"을 언급하였다. 이는 채점자 1이 "전문가적"수준의 평가 기준에서 제시된 "일관성"을 다른 수준의 평가 기준에 도입하여 재구성한 것으로 해석된다.

사례 2 : 과제1에서 예상하기를 중요한 과제 요소로 보지 않는 채점자4

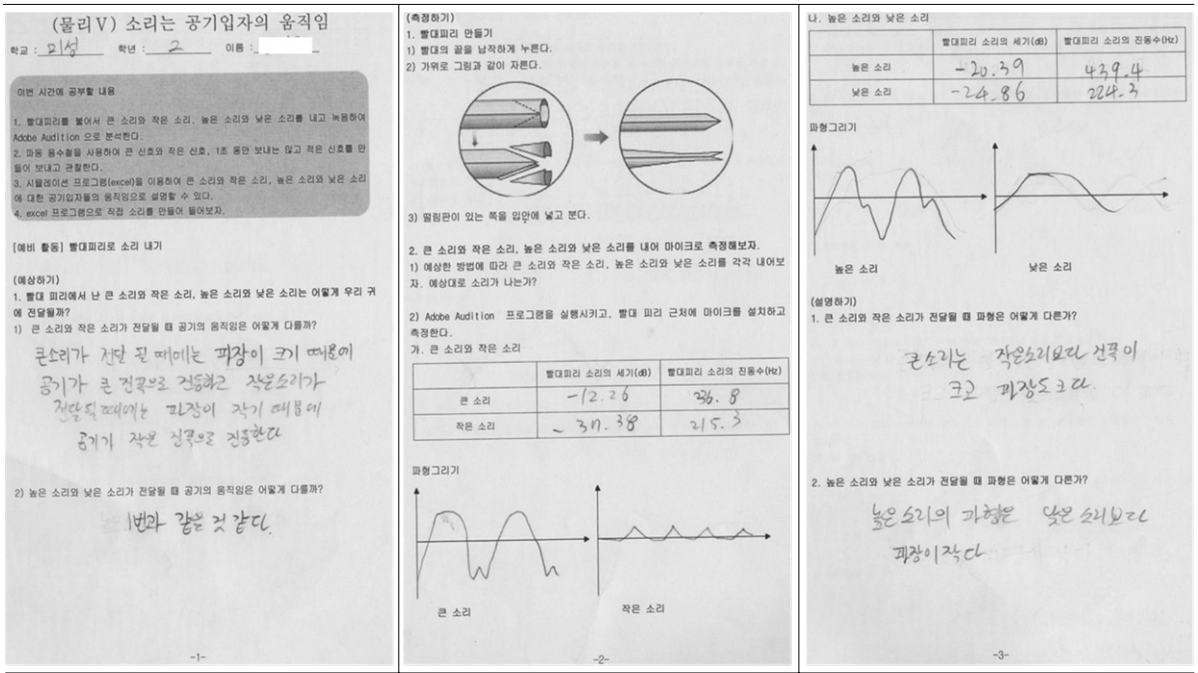
Figure 3은 과제1에 대한 학생 응답의 예시이다. 이 예시 응답에 대하여 채점자 1, 2, 3, 4는 각각 보통, 보통, 보통, 진보의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 11과 같이 설명하였다. Figure 3에 제시된 응답에서 학생은 과제를 충실히 수행하였지만, 파장이 소리의 크기에 관련된다는 예상하기에서의 생각을 설명하기에서도 나타냈다. 이 응답에 대하여 채점자 4는 예상하기를 제외한 나머지 부분에 대하여 제시된 평가 기준 중 진보 수준인 "예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다."에 적합하다고 판단하였다. 채점자 4는 이와 같이 학생의 수행 수준을 판단한 근거로 "예상하기는 이전의 개념이기 때문에 탐구활동을 진행하면서 서술한 측정하기와 설명하기가 탐구능력에 중요한 요소가 된다."와 같이 설명하였으며, 이는 채점자 4가 예상하기에서의 응답을 중요한 채점요소로 보지 않는다는 것을 나타낸다. 다른 채점자들은 예상하기의 일부만 과학적으로 타당하게 서술되었다는 이유를 들어서 이 학생의 응답이 보통 수준의 평가준거인 "예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다."에 적합하다고 판단하였다. 채점자4의 채점결과는 평가준거에 평가대상 요소가 여러 가지로 제시될 때, 채점자가 자신의 주관에 따라 중요하게 생각한 부분을 강조하면서 그렇지 않은 부분에 평가준거를 적용시키지 않았기 때문으로 해석된다.

나. 채점자가 중요하게 생각하는 평가요소가 다른 경우

채점자가 중요하게 생각하는 평가요소가 학생의 응답에 반영이 되었는지에 따라 채점자의 채점에 영향을 준 경우이다. 많은 경우에 채점자들은 자신이 중요하게 생각하는 평가요소가 답안에 포함되어 있지 않으면 다른 채점자보다 한 수준 낮게 평가하는 것으로 나타났다.

사례 3 : 과제끼리의 연계성과 그림과 설명의 연계성을 중요하게 생각한 채점자1

Figure 4는 과제 3에 대한 학생 응답의 예시이다. 이에 대하여 채점자 1, 2, 3, 4는 각각 초보, 보통, 보통, 보통의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 12와 같이 설명하였다.



과제1 예상하기	과제1 측정 및 관찰하기	과제1 설명하기

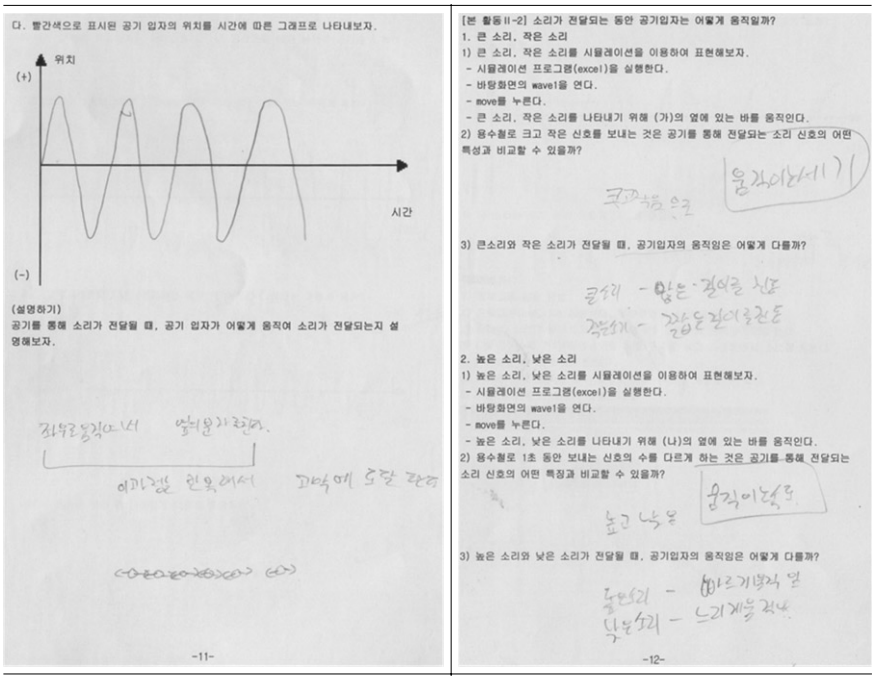
Fig. 3 *예상하기와 관찰하기 및 설명하기에서 다른 수준을 나타난 응답 예시*

Table 11
Figure 3의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자4 : 전체적으로 잘 썼다. 그런데 뒤에 파장이라는 용어로 충분히 설명하지 못해서 진보의 수준으로 채점하였다. 채점자3 : 예상하기 부분은 아예 틀렸는데. 채점자1 : 높은 소리, 낮은 소리는 아예 없다. 예상하기의 절반이 틀린 것이 된다. 그래서 보통의 수준이 맞다.
진보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자4 : 예상하기는 거의 보지 않았다. 예상하기는 이전의 개념이고 탐구활동을 진행하면서 측정하기와 설명하기를 서술하고 그린 것이 탐구능력에 중요한 요소가 된다. 채점자1 : 예상하기가 틀릴 수는 있지만 나머지 부분과 일관성이 있어야 한다. 그래야 진보 이상의 수준이라고 할 수 있다. 채점자2 : 더 낮은 수준이 맞는 것 같다. 과학적으로 옳지 않은 서술도 있다. 채점자3 : 일단 절반은 맞았으니까 중간 수준인 보통이 맞다. 채점자4 : 예상하기는 빈칸만 없어도 크게 신경 쓰지 않았다.

채점자 1은 제시된 평가준거 중 보통수준의 준거인 "... 충분히 서술되지 못했다." 보다는 초보수준의 준거인 "... 매우 부족하다."에 학생답안의 어울린다고 생각했다. 그 이유는 전체적인 답안은 진보의 채점척도에 맞았지만 "용수철로 크고 작은 신호를 보내는 것은 공기를 통해 전달되는 소리 신호의 어떤 특성과 비교할 수 있을까?"라는 질문에 학생의 응답은 "움직이는 세기"라는 애매한 응답이 제시되었다. 채점자 1은 이 질문이 앞의 용수철을 통한 과제2와 시뮬레이션을 통한 과제3의 핵심적인 연결고리라고 생각했기 때문에 제대로 서술하지 못한 응답을 고려할 때 질적으로 매우 부족하다고 판단한 것으로 해석된다.

사례 4 : 공기입자가 제자리에서 진동한다는 설명



과제3의 활동1 설명하기 | 과제3 활동2의 관찰 및 설명하기

Fig. 4 *각 평가요소에 대한 응답은 있으나 연계성은 부족한 응답 예시*

Table 12
Figure 4의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
초보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자1 : 이 학생의 수준을 낮게 준 이유는 용수철하고 소리 신호하고 비교하는 부분이 너무 약하기 때문이다. 과제 간의 연결고리가 되는 부분이기 때문에 중요하다.
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자4 : 근데 중학교 과정에서는 소리 세기로 하니까. 움직이는 높낮이도 진폭을 의미한다고 보아야 한다. 소리가 움직이고 있는데 세게 움직인다. 그래서 의미가 다 통한다. 다른 서술을 보면 알 수 있다. 채점자3 : 앞의 설명으로 유추가 가능하고 다른 부분의 서술은 중간정도 수준의 답안이라고 볼 수 있다.

을 중요하게 생각한 채점자 2

Figure 5의 학생 답안에 대하여 채점자 1, 2, 3, 4는 각각 진보, 보통, 진보, 진보의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 13과 같이 설명하였다.

채점자 2는 제시된 평가준거 중 진보수준의 준거인 "... 모두 과학적으로 타당하게..." 보다는 보통수준의 준거인 "... 일부만 과학적으로 타당하게..."에 학생답안의 어울린다고 생각했다. 채점자 2는 하나의 움직임에서 제자리에서 고정되어 앞뒤로 진동하는 종파적인 견해가 중요하다고 생각했기 때문에 이러한 서술이 포함되어 있어야 모두 과학적으로 타당하다고 본 것으로 판단했다고 해석할 수 있다.

사례 5 : 용어의 중요성보다 맥락 상 의미를 중요하게 생각한 채점자 3

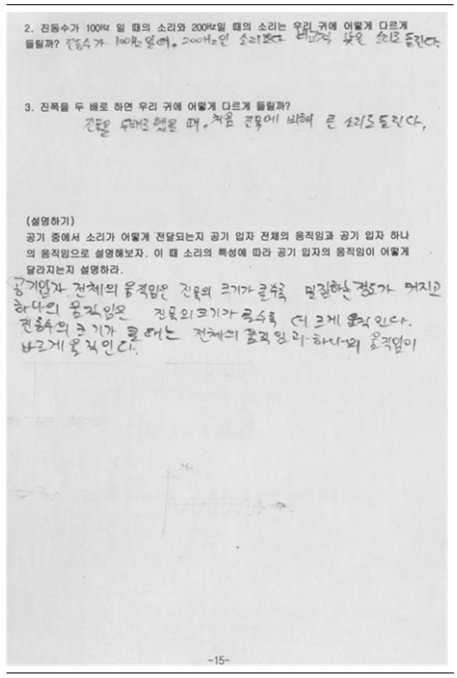
Figure 6의 학생 답안에 대하여 채점자 1, 2, 3, 4는 각각 보통, 보통, 진보, 보통의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 14와 같이 설명하였다.

채점자 3은 제시된 평가준거 중 진보수준의 준거인 "... 모두 과학적으로 타당하게..."에 학생답안의 어울린다고 생각했다. 학생의 용어 중 파장이라는 용어가 잘못 사용되었지만 맥락상 의미가 통한다고 생각하여 과학적으로 타당한 서술로 보았다. 이는 학생의 답안이 애매한 경우 맥락상 의미가 통하면 용어의 사용에 의미를 두지 않는 채점자와 의미를 두는 채점자에 의해 서로 다른 채점을 하게 된다고 해석할 수 있다.

진보, 보통, 초보 수준의 구별이 어려운 것은 학생응답의 질과 양의 정도를 구분하는 부분이었고 특히 채점자별로 중요하게 생각하는 요소가 다르면 구분하는 판단의 중요도에 따라 달라져서 한 수준이나 두 수준 정도의 차이로 나타났다.

다. 채점자별로 중요하게 생각하는 요소와 엄격성의 경향

Table 15는 채점자별로 중요하게 생각하는 요소를 채점자간 협의회를 진행한 후 면담하여 정리한 것이다. 본 연구에서 제시한 네 개의 탐구활동 과제는 각


2. 진동수가 100Hz 일 때의 소리와 200Hz일 때의 소리는 우리 귀에 어떻게 다르게 들릴까? 진동수가 100Hz일때가 200Hz인 소리보다 비교적 낮은 소리로 들린다.

3. 진폭을 두 배로 하면 우리 귀에 어떻게 다르게 들릴까?
진폭을 두배로 했을 때, 처음 진동에 비해 큰 소리로 들린다.

(설명하기)
공기 중에서 소리가 어떻게 전달되는지 공기 입자 전체의 움직임과 공기 입자 하나의 움직임으로 설명해보자. 이 때 소리의 특성에 따라 공기 입자의 움직임이 어떻게 달라지는지 설명하라.
공기입자 전체의 움직임은 진폭의 크기가 클수록, 밀집한 정도가 커지고 하나의 움직임은 진폭의 크기가 클수록 더 크게 움직인다. 진동수의 크기가 클때는 전체의 움직임과 하나의 움직임이 비교적 작게 움직인다.

-15-

과제4 설명하기

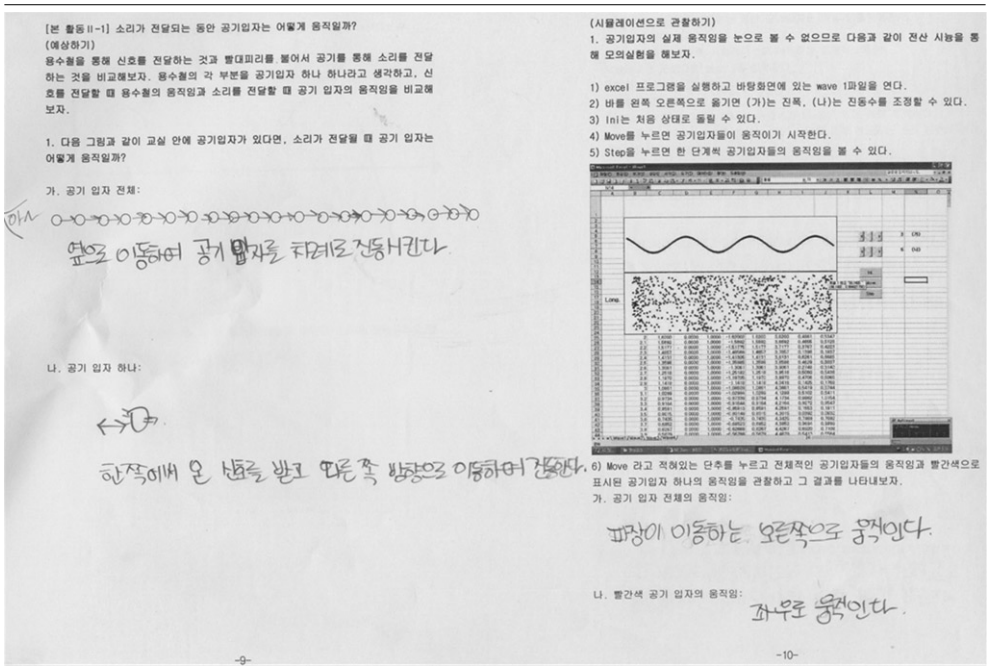
Fig. 5 *채점자2이 중요하게 생각하는 평가요소가 누락된 응답 예시*

Table 13
Figure 5의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자2 : 단순히 움직임이 커지거나 빨라진다고 서술하는 것은 제자리에서 진동한다는 의미를 담아내지 못한다. 최소한 진동이라는 말을 써야만 그 의미가 전달이 된다. 채점자3 : 완전하지는 않지만 의미 전달은 되고 틀렸다고 보기는 어렵다. 채점자4 : 서술이 불완전하기는 하지만 과학적으로 타당하다고 유추할 수 있다.
진보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자2 : 설명하기에서 공기입자 하나의 움직임이 명확하게 서술되어야 과학적으로 타당하게 서술했다고 할 수 있다.

각 예상하기, 관찰/측정하기, 설명하기로 구성되어 있다. 총체적 채점 방식을 선택하였기 때문에 채점 준거는 채점 대상과 평가요소를 분석적으로 명시하지 않고, "예상하기, 측정하기, 설명하기의 각 부분의 설명이 (일관되고 모두 과학적으로 타당하게/일부만 과학적으로 타당하게/대부분 과학적으로 타당하지 않게) 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 (충분히 서술되었다/충분히 서술되지 못했다/매우 부족하다."와 같이 포괄적으로 제시하였다. 이러한 채점 기준에 대하여 채점자들은 채점 대상이 되는 과제 요소인 예상하기, 측정하기, 설명하기에 대하여 각기 다르게 가중치를 배분한 것으로 해석된다. 또한 "일관된 설명"에 대하여 "과제 사이의 연계성"이나 "그림과 설명 사이의 연계성"으로 해석하는 경우와 "공기입자의 움직임에 대한 충분한 설명"에 대해서는 "용어보다는 맥락 상의 의미를 중요"하다는 해석도 나타났다.

선행 연구결과에 의하면, 채점 대상의 요소를 명시하는 분석적 채점의 경우 채점자간 일치도가 총체적 채점보다 높게 나타나며, 총체적 채점의 경우는 채점



과제3의 활동 1 예상하기 | 과제3의 활동1 관찰하기

Fig. 6 *맥락이 잘 드러나지 않는 학생 응답 예시*

Table 14
Figure 6의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자1 : 파장과 파동은 다르다. 채점자2 : 앞 뒤로 맥락을 따지면 파장을 파동으로 해석해도 큰 문제는 없지만 공기입자 하나가 좌우로 움직인다는 것은 제자리에서 진동한다는 의미가 없다.
진보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자3 : 전체가 오른쪽으로 움직인다는 뜻으로 보면 파장이 결국 파동을 의미하므로 별 문제가 없다고 생각한다.

Table 15
채점자별 엄격성의 경향과 중요하게 생각하는 요소

엄격성	채점자	채점자가 중요하게 생각하는 요소	내적 일치도	채점자간 일치도
엄격한 채점자	채점자1	과제끼리의 연계성과 그림과 설명의 연계성을 중요하게 생각함	0.77	.92*
	채점자2	공기입자가 제자리에서 진동한다는 설명을 중요하게 생각함	0.81	
관대한 채점자	채점자3	용어의 중요성보다 맥락상의 의미를 중요하게 생각함	0.74	.89*
	채점자4	과학적 용어가 부정확하다해도 문맥상 대략적인 의미를 중요하게 생각함	0.81	

*p<.01

자의 내적 일치도가 높게 나타나는 것으로 보고되었다(김형준, 2010). 이는 총체적 채점 준거에서 채점 대상별 평가요소를 평가준거에서 명시하지 않기 때문에 나타나는 현상으로 해석할 수 있다. 즉 총체적 채점에서 채점자마다 중요시하는 대상과 요소가 달라서 채점자간 일치도가 분석적 채점보다는 낮아질 수 있지만, 채점자 자신이 중요시하는 요소를 채점자 스스로 명료화하기 때문에 채점자 내적 일치도가 높게 나타날 수 있다고 해석할 수 있다.

동일한 채점표에 대하여 채점자 1과 채점자 2는 여러 가지 채점 요소 중 가장 중요한 요소를 하나 정하고, 학생의 응답에서 해당 요소가 명시적으로 포함되었는지의 여부로 점수를 준 결과 좀 더 엄격한 채점자로 나타났다고 해석할 수 있다. 용어보다는 설명의 맥락상 의미를 중요시하한다고 응답한 채점자 3과 채점자 4는 학생의 응답에서 명시적으로 드러나지 않는 문맥상의 숨은 의미를 해석하여 점수를 준 결과 채점자 1과 채점자 2보다 관대한 채점자로 나타난 것으로 해석할 수 있다. 채점자 1과 채점자 2는 총체적 채점 기준표를 보고 분석적으로 채점한 것으로 해석할 수 있으며, 채점자 3과 채점자 4는 총체적 채점의 의미를 살려 맥락상의 의미를 채점한 것으로 해석할 수 있다. 그러나 맥락상의 의미를 채점한다는 것이 무엇을 의미하는 가에 대한 논의는 계속해서 필요하다고 할 수 있다.

3. 불일치 유형2: 엄격함과 관대함의 경향에 의한 불일치

채점자간 불일치 유형2는 채점자의 엄격함과 관대함의 경향에 의해 불일치가 발생하는 경우이다. 총 240사례 중 47사례가 불일치 유형2로 인식되었다. 이 불일치 유형이 발생하는 경우는 대부분 학생 응답의 수준이 채점 척도의 경계에 있는 경우로 채점자의 엄격함과 관대함의 경향에 따라 관대한 채점자는 높은 점수를, 엄격한 채점자는 낮은 점수를 주어서 발생한다. Table 16에 따르면, 이 불일치 유형이 발생한 47사례 중 44사례(93%)가 1수준의 차이를 나타내며, 주로 초보와 보통, 또는 보통과 진보 수준을 판단하는데 발생하는 것으로 나타난다.

사례 6 : 초보와 보통 사이에서 엄격한 채점자 1, 2와 관대한 채점자 3, 4

Figure 7의 학생 답안에 대하여 채점자 1, 2, 3, 4는 각각 초보, 초보, 보통, 보통의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 17과 같이 설명하였다. 엄격한 채점 경향을 보이면서 과제 사이의 연계 혹은 공기입자의 움직임을 중요한 채점 요소로 해석한 채점자 1과 채점자 2는 “공기입자의 움직임을 입자가 세게 부딪친다고 했는데, 이것으로는 설명이 부족하다”고 Figure 7의 학생 응답을 초보로 채점한 이유를 설명하였다. 동일한 응답에 대하

Table 16
불일치 유형2에 의한 채점 결과의 차이

	채점척도 차이	과제1	과제2	과제3	과제4	계	
1수준 차이	전문가적 ↔ 진보	5	1	1	1	8	44
	진보 ↔ 보통	1	1	3	5	10	
	보통 ↔ 초보	1		7	7	15	
	초보 ↔ 설명못함	5	3		3	11	
2수준 차이	전문가적 ↔ 보통	1			1	2	3
	진보 ↔ 초보					0	
	보통 ↔ 설명못함			1		1	
3수준이상 차이	전문가적 ↔ 초보					0	0
	진보 ↔ 설명못함					0	
	전문가적 ↔ 설명못함					0	
계		13	5	12	17	47	

여 관대한 채점자이면서 맥락상 의미를 중요한 채점 요소를 판단한 채점자3과 채점4는 "..의미가 통하면 몇 개 틀려도 전반적으로 후하게 채점하였다' 라고 Figure 7의 학생 응답을 보통으로 채점한 이유를 설명하였다.

제시된 평가준거 중 보통수준의 준거인 "...충분히 서술되지 못했다."와 초보수준의 준거인 "...설명이 매우 부족하다"등과 같은 정성적인 기술은 학생의 응답이 보통이나 초보 수준에 명시적으로 부합하는 경우에는 채점자들이 큰 불일치가 없이 학생의 응답을 변별할 수 있다. 그러나 학생 응답의 수준이 보통과 초보의 경계에 놓이게 되면 채점자들은 자신의 판단을 담보하기 위하여 분석적 기준으로 도입하여 학생의 응답을 평가하는 것으로 해석할 수 있다. 즉, 학생 응답을 변별하기 위하여 학생의 응답 중 과학적으로 옳은 부분이 전체 중 몇 개가 되는가와 같은 수량화를 하는데, 이때 채점자마다 각기 다른 기준을 적용하며 이 기준의 애매함 때문에 관대한 채점자와 엄격한 채

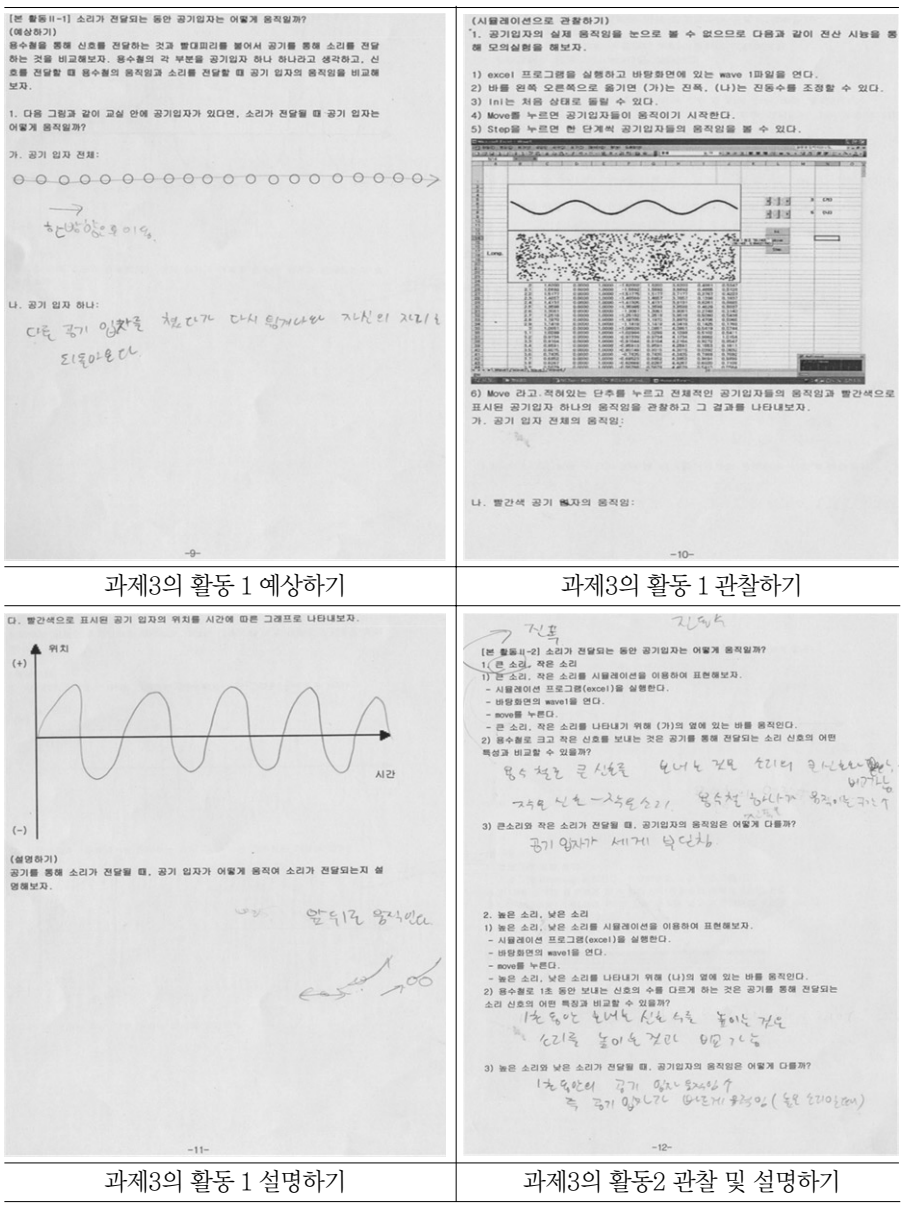


Fig. 7 *초보와 보통 사이의 수준에 있는 응답 예시*

Table 17
Figure 7의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
초보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자1 : 큰소리 작은 소리에서 공기입자의 움직임을 입자가 세게 부딪친다고 했는데 이것으로는 설명이 부족하다. 또 관찰하기 부분에서 아예 안 쓴 부분이 있다. 그래서 초보 이상의 수준을 주기에는 어렵다. 채점자2: 채점자1과 비슷하다.
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자3 : 과제 3을 전반적으로 후하게 수준을 부여한 것 같다. 채점자4 : 과학적으로 타당한 것과 서술되는 양에서 애매함이 있다. 그 경우 의미가 통하면 몇 개 틀려도 전반적으로 후하게 채점하였다.

점자 사이의 불일치가 발생한다고 해석할 수 있다.

사례 7 : 보통과 진보사이에서 엄격한 채점자 1, 2와 관대한 채점자 3, 4

Figure 8의 학생 답안에 대하여 채점자 1, 2, 3, 4는 각각 보통, 보통, 진보, 진보의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 18과 같이 설명하였다. 채점자 1, 채점자 3, 채점자 4가 모두 Figure 8의 응답이 보통과 진보 수준의 중간이라고 판단하였다. 그러나 부여한 수준은 관대한 경향의 채점자 3과 채점자 4는 진보, 엄격한 채점자 1과 채점자 2는 보통으로 나타났다. 채점자들은 총체적 채점 준거 중 "일부"란 전체 중 몇 개인가와 "모두 과학적으로 타당한 서술"은 몇 개인가 등과 같이 정성적으로 제시된 평가 준거를 수량적으로 해석하게 되었다고 설명하였다.

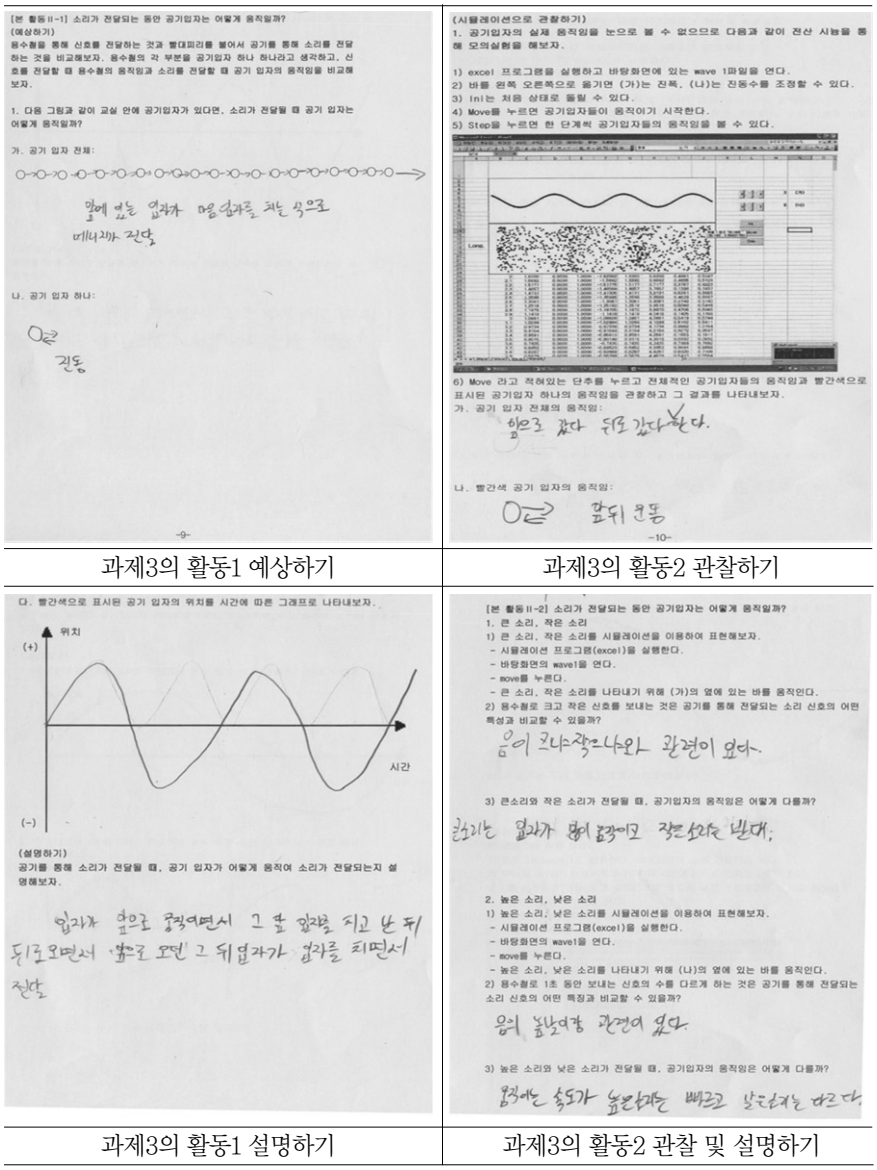
두 사례에 대한 면담 결과를 종합해보면 학생의 응답이 두 수준 사이의 경계에 있는 경우에 맞는 서술과 틀린 서술의 개수에 따라 채점자의 판단이 결정짓는 것으로 해석할 수 있다. 사례 6의 경우 학생의 응답에서 맞는 서술은 5개, 틀린 서술이 2개였다. 사례 7의 경우, 맞는 서술은 9개, 틀린 서술이 0개였다. 이들 사례가 경계에 있는지 확인하기 위하여 과제 3의 12사례에 대하여 채점자별로 부여한 수준에 따른 맞은 서술과 틀린 서술의 개수를 조사하여 Table 19에 제시하였다.

Table 19의 결과에 따르면, 총체적 채점방식에서 채점자들에게 명시적으로 맞은 서술과 틀린 서술의 개수를 세는 것을 요구하지 않았지만 채점자들이 부여한 수준에 따라 학생 응답 중 맞은 서술과 틀린 서술의 개수가 일정한 양상을 나타냈다. 즉, 엄격한 채점자인 채점자 1, 2의 기준은 관대한 채점자인 채점자 3, 4보다 1~2개 이상 더 많이 나타났다. 보통과 초보의 수준을 변별하는 기준은 전문자적과 진보 수준을 변별하는 기준보다 더 차이가 많이 나타났다. 이는 채점자들이 맞은 서술의 개수 혹은 틀린 서술의 개수에 따라 학생 응답의 수준을 판단한 것으로 해석할 수 있다. 사례 6과 사례 7은 맞는 서술의 개수가 5개와 7개로 진보와 보통, 보통과 초보사이의 경계에 놓인 응답이다. 이러한 응답을 변별하기 위하여 채점자들이 제시된 총체적인 평가준거로부터 자체적으로 학생의 응답 수준을 판단하는 양적 기준을 설정한다고 해석할 수 있다.

이와 같은 한 채점자의 관대함과 엄격함의 경향이 일관성있게 나타나는 불일치는 통계적으로 처리가 가능하다. 고전적으로 평균, 표준점수의 방법으로 보정하기도 하지만, 최근에는 일반화가능도이론을 응용하여 평가자의 관대함과 엄격함의 경향을 엄격성 모수치로 추정하고 이를 반영하여 학생의 능력 추정치를 산출하고 있으며, 채점자의 엄격성 모수치가 다양한 경우에는 특히 다국면 라쉬 모형에 의한 보정이 효과적인 것으로 알려졌다(설현수, 2010; 지은림, 2008).

4. 불일치 유형3: 한 채점자가 중요한 요소를 간과해서 발생한 불일치

불일치 유형3은 한 채점자가 중요한 과제요소나 평가요소를 간과하여 발생하는 불일치유형이다. Table 20에 따르면 채점자가 중요한 요소를 간과해서 발생



과제3의 활동1 예상하기	과제3의 활동2 관찰하기
과제3의 활동1 설명하기	과제3의 활동2 관찰 및 설명하기

Fig. 8 *보통과 진보 사이의 수준에 있는 응답 예시*

Table 18
Figure 8의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
보통	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자3 : 관대한 사람과 엄격한 사람의 차이가 잘 드러나는 학생 답안이다. 채점자1 : 몇 점에서 차이가 났지? 중간점수인가? 채점자4 : 보통에서 진보사이이다. 이런 경우가 많다. 채점자1 : 중간이 채점하기가 애매하기 때문이다. 채점자4 : 경계에서 관대함과 엄격함의 차이로 발생했다. 보통하고 진보사이의 경계가 애매하다. 이런 경우에 나와 채점자3은 높은 수준으로 결정했다.
진보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	

Table 19
채점자 경향에 따른 맞은 서술과 틀린 서술 개수 (과제 3)

채점자경향		수준				
		전문가적	진보	보통	초보	설명못함
엄격한 채점자 (채점자1,2)	맞은 서술	12	10~11	7~9	5~6	0~4
	틀린 서술	0	0	0~2	2~5	6이상
관대한 채점자 (채점자3,4)	맞은 서술	12~11	9~10	5~8	3~4	0~3
	틀린 서술	0	0	1~2	3~6	7이상

Table 20
불일치 유형3에 의한 채점결과의 차이

	채점결과의 차이	과제1	과제2	과제3	과제4	계	
1수준 차이	전문가적 ↔ 진보	2	1		3	6	27
	진보 ↔ 보통	1	1		5	7	
	보통 ↔ 초보		4	4	1	9	
	초보 ↔ 설명못함		1	4		5	
2수준 차이	전문가적 ↔ 보통	2	2	2	3	9	26
	진보 ↔ 초보	5	2	4	1	12	
	보통 ↔ 설명못함	1	1	2	1	5	
3수준 이상 차이	전문가적 ↔ 초보		1				4
	진보 ↔ 설명못함		1	1		2	
	전문가적 ↔ 설명못함	1			1	2	
계		12	14	17	15	58	

하는 불일치는 240사례 중 58사례이며, 불일치의 정도가 2~3수준되는 경우가 불일치 유형1과 2에 비해 더 많이 발생함을 볼 수 있다. 이 유형에서는 많은 경우 한 채점자가 채점 중에는 보지 못하고 간과했던 것을 다른 채점자와의 협의를 통해 알게 되고 채점을 바로 잡아 자신의 실수로 인정하게 된다. 특히 불일치의 정도가 클수록 이런 경우가 많이 발생하고 있다. 이 불일치 유형은 채점자 자신이 중요하게 생각하지 않는 부분에 대해서는 쉽게 지나치는 경향으로 발생한다고 해석할 수 있다.

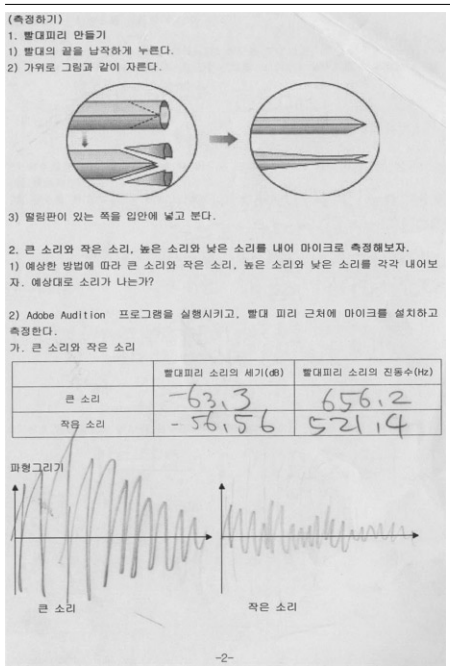
사례 8 : 중요 요소를 간과해서 오차가 발생한 사례

Figure 9의 학생 답안에 대하여 채점자 1, 2, 3, 4는 각각 초보, 초보, 진보, 초보의 수준으로 판단하였다. 각각의 채점자들은 자신의 판단에 대하여 Table 21과 같이 설명하였다. 이 경우는 채점자 3이 측정하기의 그래프를 소홀히 본 결과 동일한 학생의 응답에 대하여 다른 채점자들보다 두 수준 위인 진보를 판단하였다. 이는 채점자가 자신이 중요시하는 요소는 아니지만, 응답의 수준을 결정하는 중요한 요소인 그래프나 그림을 간과하였기 때문에 학생 응답의 질을 제대로 판단하지 못하여 나타난 사례이다.

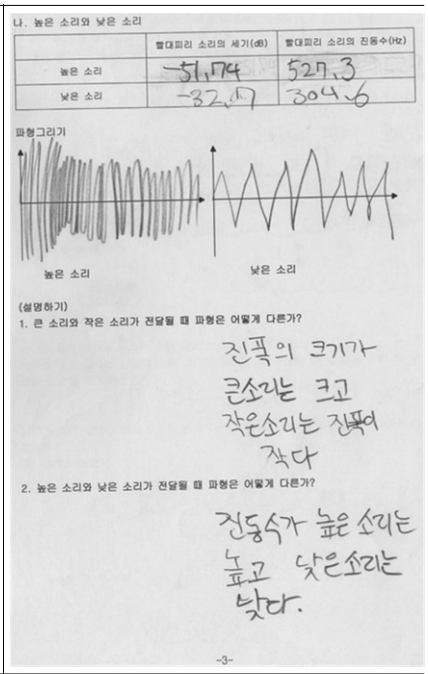
이와 같은 불일치 유형이 나오지 않는 것이 바람직하나 현실적으로 발생하게 된다. 이와 같은 불일치가 발생하는 것을 방지하기 위해서는 2인 이상의 채점자가 교차 채점하는 것이 필요하며, 채점자간 불일치가 큰 사례에 대해서 재평가를 실시하는 것이 바람직하다.

Ⅲ. 결론 및 시사점

본 연구는 중학교 과학탐구활동 수행평가 채점시 나타나는 채점자간 불일치의 정도와 유형을 분석하기



과제1 관찰하기



과제1 설명하기

Fig. 9 *채점자가 중요요소를 간과한 사례*

Table 21
Figure 9의 응답 예시 채점 결과에 대한 채점자의 설명

평가 결과	평가 준거	채점자의 설명
초보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일부만 과학적으로 타당하게 서술되었고 소리의 특성에 따라 공기입자의 움직임에 대한 설명이 매우 부족하다.	채점자3 : 아! 설명하기를 집중해 보느라 측정하기의 그래프를 소홀히 보았다. 내 실수다. 이건 초보로 판단해야 한다.
진보	예상하기, 측정하기, 설명하기의 각 부분의 설명이 일관되고 모두 과학적으로 타당하게 서술되었지만 소리의 특성에 따라 공기입자의 움직임이 충분히 서술되지 못했다.	채점자1 : 측정하기를 보면 파형에 높은 소리, 낮은 소리, 큰소리, 작은 소리가 잘 드러나지 않는다. 채점자2 : 실제 Adobe Audition에 나타난 그래프와 차이가 많다. 채점자4 : 그래서 초보로 판단했다.

위하여 60명의 중학생을 대상으로 과학탐구 수행평가를 실시하고 학생의 응답을 수합하여 4명의 채점자가 채점을 하였다. 채점자 관련 오차가 채점결과에 미치는 영향을 파악하기 위하여 분산성분별 분산 추정치, 점수의 표준화를 통한 채점자별 관대함의 경향, 불일치의 빈도수 등을 조사하였다. 또한 채점자간 불일치가 발생하는 사례에 대해서는 채점자 협의회를 통하여 채점자간 불일치 유형을 합의하고 분류하였다. 일반화가능도 이론에 의한 분산 분석 결과 학생 분산 성분으로 전체 분산의 47%를 설명할 수 있었으며, 채점자, 채점자와 학생의 상호작용, 채점자와 과제의 상호적용 등 채점자 관련 분산성분으로 전체 분산의 25%를 추정할 수 있었다. 또한 4명의 채점자 중 2명은 관대한 채점자, 다른 2명은 엄격한 채점자의 경향을 나타냈으며, 총 240사례 중 4명의 채점 결과가 모두 일치하는 사례는 51사례였다. 나머지 189사례에 대하여 채점자 간 협의를 통하여 확인한 결과 채점자들이 인식한 채점자간 불일치 유형은 크게 세 가지로 나타났다.

채점자간 불일치 유형1은 채점자가 중요하게 생각

하는 부분이 다르기 때문에 발생하는 불일치로 채점자마다 중요하게 생각하는 과제 요소와 평가 요소가 다르기 때문에 발생한다. 사례분석을 통해 채점자들이 제시된 총체적 채점기준표의 평가 기준을 해석할 때, 맥락상의 의미를 강조하여 해석하거나 용어 등 특정한 요소만을 강조하여 분석적으로 해석하는 경향에 의한 것으로 해석할 수 있다. 일반화 가능도 이론에 의해 엄격한 채점자로 분류된 채점자1과 채점자2는 보다 분석적인 경향을 나타냈으며, 관대한 채점자로 분류된 채점자 3과 채점자 4는 보다 맥락을 중시하는 경향을 나타냈다. 불일치 유형1을 줄이기 위해서는 출제자와 채점자, 채점자와 채점자 사이에 주어진 총체적 평가 기준에 대한 이해와 해석이 공유되어야 한다. 이를 위해서는 출제자와 채점자간 협의가 충분해야 하며, 특히 수행평가 과제가 여러 요소로 구성된 경우 중요하게 생각해야 할 과제 요소가 어떤 것인지를 서로 협의하는 것이 불일치를 줄이는 데 도움을 줄 것이다. 또한 채점자들이 각자 선호하는 평가 요소에 관련된 응답이 포함되는 경우나 포함되지 않는 경우에 한 수준 정도의 차이로 판단이 달라지기 때문에 이 부분 역시 채점자간 협의를 통해 조율하는 것이 필요하다고 제안한다. 동일한 채점 기준에 대한 채점자의 해석이 다르게 나타나는 경향과 원인에 대한 추후 연구가 필요하다.

불일치 유형2는 채점자의 엄격함과 관대함의 차이로 인해 발생하는 것으로 전체 불일치 사례 중 25%로 나타났다. 채점자의 엄격함과 관대함에 의해 발생하는 불일치 유형2에 대한 사례분석을 통해 채점척도의 경계에 놓인 학생의 응답을 채점할 때는 채점자가 총체적 평가 준거를 분석적이고 수량화하는 경향을 확인할 수 있었다. 학생의 응답 중 옳은 서술의 개수에 따라 학생의 응답 수준을 결정하는 경향을 나타냈는데, 채점자의 엄격함과 관대함의 경향에 따라 각 수준에서 해당하는 옳은 서술의 개수가 달라졌다. 이 불일치 유형을 줄이기 위해서는 총체적 채점 방식을 선택한 경우라도 각 수준에 대한 평가 기준과 함께 수준의 경계에 대한 기술을 명확히 할 필요가 있다. 또한 채점자들은 자신의 채점경향이 다른 채점자에 비하여 엄격한지 관대한지를 사전에 파악하고 경계에 놓인 학생들의 응답에 대한 판단을 주의 깊게 하여야 할 것으로 판단된다. 관대함과 엄격함의 경향이 일관성있게 나타나는 경우는 다국면 라쉬 모형으로 보정할 수도 있다.

불일치 유형3은 채점자의 실수로 발생하는 불일치로 전체 불일치 사례 중 31%로 나타났다. 본 연구에서 나타난 채점자의 실수는 채점자가 본인이 중요시하는 요소외의 다른 평가 요소나 과제 요소를 간과해서 발생하는 경우이다. 이 불일치 유형에서는 채점자간 채점 결과의 차이가 크게 나타나기 때문에 점수의 오차가 커질 수 있다. 채점자간 불일치 유형3은 채점자 혼자서 채점할 때는 드러나지 않는 문제이기 때문에 여러 명의 채점자가 교차 채점을 하고 채점결과가 큰 사례에 대한 재평가를 통해 불일치를 줄일 수 있다.

본 연구 결과에 의해 구체적으로 예시된 채점자의 채점자 관련 불일치 유형은 실제 채점에서 채점자 훈련에 활용하여 채점의 신뢰도를 높이는데 사용될 수 있다. 동일한 채점 기준에 대하여 채점자마다 다르게 해석하는 경향과 원인에 대한 추후 연구가 필요하다.

국문 요약

본 연구의 목적은 과학탐구활동 수행평가시 총체적 채점의 신뢰도를 높이기 위하여 채점자간 불일치의 정도와 유형을 이해하는 것이다. 이를 위하여 중학생 60명을 대상으로 과학탐구 수행평가를 실시하였고, 4명의 훈련된 채점자 채점을 실시하였다. 분산 분석결과 교사 관련 분산성분에 의해 전체 분산의 25%를 설명할 수 있으며, 4명의 채점자중 2명은 관대한 채점자, 2명은 엄격한 채점자의 경향을 지닌 것으로 나타났다. 전체 240 채점 사례 중 4명의 채점자가 모두 일치한 사례는 51사례이다. 채점자간 불일치가 나타나는 189사례에 대하여 채점자 협의를 통하여 확인한 결과, 채점자간 중요하게 생각하는 부분의 차이 때문에 발생하는 불일치 유형1이 38%, 채점자의 관대함과 엄격함에 의해 발생하는 불일치 유형2가 25%, 채점자가 중요하게 생각하지 않은 부분에서 간과하기 때문에 발생하는 등 실수에 의한 불일치 유형3이 31%로 나타났다. 불일치 유형1은 채점자마다 중요하게 생각하는 과제 요소와 평가 요소가 다른 경우로 나누어서 나타났으며, 맥락상의 의미를 강조하는 채점자는 관대한 경향을, 특정 요소를 강조하여 분석적으로 해석하는 채점자는 엄격한 경향을 나타냈다. 불일치 유형2는 많은 경우 채점 척도의 경계에 학생의 응답에 대

하여 나타났으며, 채점자들은 이러한 학생 응답에 대하여 옳은 서술의 개수를 세는 등 분석적인 채점을 수행하는 것을 확인할 수 있었다. 또한 불일치 유형3은 채점자의 실수로 발생하는 불일치로 주로 학생의 응답 중 평가 기준에 부합하는 부분인데 채점자가 중요하게 생각하지 않기 때문에 간과하여 발생하는 것으로 파악할 수 있었다. 이상과 같은 채점자간 불일치를 제어하기 위해서는 채점자가 중요하게 생각하는 과제 요소와 평가 요소에 대하여 사전 및 진행 중 협의를 할 필요가 있다고 판단된다. 또한 총체적 채점을 하는 경우도 각 수준에 해당하는 평가 기준과 함께 경계에 놓인 학생 응답을 변별하는 기준을 제시하는 것이 필요하다. 채점자들은 자신의 채점경향이 엄격한지 관대한지를 파악하고 경계에 놓인 학생의 응답에 대한 판단을 주의 깊게 하여야 불일치를 줄일 수 있다. 실수에 의한 오차를 줄이기 위해서는 여러 명의 채점자가 교차 채점하는 것이 필요하다. 동일한 채점 기준에 대한 채점자의 해석이 다르게 나타나는 경향과 원인에 대한 추후 연구가 필요하다.

참고 문헌

김명숙(1999). 영어작문 수행평가의 채점행위 분석 연구. 교육평가연구, 12(2), 25-54.

김형준, 유준희(2010). 중학생 과학탐구활동 수행평가 시 채점 방식 및 척도의 수에 따른 신뢰도 분석, 한국과학교육학회지, 30(2), 275-290.

설현수(2010). 평정자간의 엄격성 차이 정도가 피험자 총점산출 방법에 미치는 영향: 원점수, 표준점수, Facet점수 비교. 교육평가연구, 23(1), 125-147.

성태제(2005). 문항반응이론의 이해와 적용. 교육과학사.

송미영, 김수진, 김희경, 남명호(2009). 온라인 시스템을 활용한 대규모 서답형 평가의 채점 일관성. 교육평가연구, 22(3), 827-846.

이규민(2007). 초등학교 과학과 수행평가의 총체적 채점과 분석적 채점 방식에 대한 일반화가능도분석. 아동교육, 16(4), 169-184.

지은림(1999). 사회과 보고서 수행평가를 위한 총체적 채점과 분석적 채점의 비교. 교육평가연구, 12(2), 11-24.

지은림(2008). 논술고사의 신뢰성에 영향을 미치는 채점자 특성 분석. 교육평가연구, 21(2), 97-113.

Black, P. J. (1990). APU science - the past and the future. School Science Review, 72(258), 28-43.

Clauser, B. E. Clayma, S. G., & Swanson, D. B. (1999). Components of rator error in a complex performance assessment. Journal of Education Measurement, 36(1), 29-45.

Clauser, B. E., Harik, P., & Margolos, M. J. (2006). A multivariate generalization analysis of data from performance assessment of physicians' clinical skills. Journal of Educational Measuremnet, 43(3), 173-191.

Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., Rosengrant, D., & Warren, A. (2006). Developing and assessing student scientific abilities. Physical Review Special Topics - Physics Education Research, 2(2), 020103-1-020103-15.

Guilford, J. P.(1954). Psychometirc Methods. Mcgraw-Hill.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. International Journal of Science Education, 25(12), 1509-1528.

Halonen, J. S., Bosack, T., Clay, S., & McCarthy, M. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. Teaching of Psychology, 30(3), 196-208.

Harick, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., & Swanson, D. (2009). AN examination rater drift within a generalizability theory framework. Jopurnal of Education Measurement, 46(1), 43-58.

Klein, S. P., Stecher, B. M., Shavelson, R., McCaffrey, D., Bell, R. M., Comfort, K., Othman, A. R., & Ormseth, T. (1998). Analytic versus holistic scoring of science performance tasks. Applied Measurement in Education, 11(2), 121-137.

Myford, C. & Wolfe, E. (2003). Detecting and measuring rater effects using many facet rasch measurement: Part 1. Journal of Applied Measurement, 4(4), 386-422.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research, Language Learning, 47(1), 101-143.

Waltman, K., Kahn, A., & Koency, G. (1998). Alternative approaches to scoring: The effects of using different scoring methods on the validity of scores from a performance assessment. CSE Technical Report, 488.

Wilson, M. & Case, H. (1997). An examination of variation in rater severity over time: A study in rater drift. BEAR report, University of California, Berkeley.

Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied Measurement in Education, 13(2), 181-208.

Woolnough, B. E. (1989). Toward holistic view of precesses in science education, in J. Wellington (ed.) Skills and processes in science education: a critical analysis. Routledge.