교육학석사학위논문

차등배점 조건을 고려한 학력검사의 신뢰도 분석

- 일반화가능도 이론을 중심으로 -

2016 년 8 월

서울대학교 대학원

교육학과 교육학 전공

이 신 혜

차등배점 조건을 고려한 학력검사의 신뢰도 분석

- 일반화가능도 이론을 중심으로 -

지도교수 박 현 정

이 논문을 교육학석사 학위논문으로 제출함

2016 년 **8** 월

서울대학교 대학원

교육학과 교육학 전공

이 신 혜

이신혜의 석사 학위논문을 인준함

2016 년 **8** 월

위 원 장 백 순 근 (인)

부위원장 신 종 호 (인)

위 원 박 현 정 (인)

국문초록

우리나라의 중·고등학교에서 주로 쓰이는 검사는 문항 유형과 배점의 다양화를 통해 검사의 변별력을 높이는 혼합형 검사이다. 이는 등급이 부여되거나 선발이 필요한 성적관리 체계에서 동점자가 생기는 것을 방지하고, 문항의 수준과 풀이 시간 등을 고려하여 검사의 타당성을 높이기 위함이다. 그러나 교육측정학적 측면에서, 검사의 구성 요소에 따라 그 검사의 신뢰도가 달라질 수 있음을 충분히 고려하지 않는다면 다양한 유형의 문항과 배점방식은 오히려 신뢰도를 떨어뜨려 검사의 타당성마저 위협할 수 있다.

검사의 신뢰도 및 타당도와 관련하여 차등배점에 대한 논의는 선행연구에서 지속적으로 언급되었다. 학교 현장에서 차등배점은 변별을 목적으로 사용되는 경우가 많은데, 문항에 매겨지는 이러한 가중치는 주로 교과 전문가의 선험적 지식과 관례 또는 전문가의 판단에 의한 문항의 난이도와 중요도에 따라 결정된다. 이러한 전문가 판단 방법은 전문가가 문항과 피험자의 특성을 충실히 파악하여 판단할 경우, 검사의 활용목적에 적합한 검사 결과를 제공하고 피험자의 능력을 보다 잘 변별할 수 있다는 장점이 있다. 그러나 판단의 준거가 명확하지 않을 경우 주관적이고 임의적이라는 비판을 받을 수 있으며, 이러한 이유로 측정학적으로 타당한지에 대하여 선행연구에서 논란이 제기되어 왔다.

이 연구에서는 동등배점과 차등배점의 여러 조건을 포함한 배점 조건을 비교하여 각 조건에 따라 검사의 신뢰도가 어떻게 달라지는지 알아보고, 피험자 분포나 문항 수의 변화에 따라서도 신뢰도를 비교하여 분석하고자 하였다. 이를 위해 일반화가능도 이론을 중심으로 모의자료를 생

성하여 검사의 신뢰도를 분석하였다. 모의자료는 최근 8년간의 대학수학능력시험 수리영역의 문항 구성을 참고하였으며, 생성조건은 총 4가지로 구성하여 분석하였다. 우선, 동등배점 조건과 차등배점 조건 3가지를 고려하여 배점 조건을 구성하였고, 그 다음으로 개별 학교에서의 검사임을 감안하여 피험자 규모를 100명, 300명으로 설정하여 모의자료를 생성하였다. 피험자 분포가 정규분포를 따를 때와 부적편포를 따를 때로 나누어 피험자 분포를 생성하였고, 이후 이를 일반화가능도 이론을 적용하여 500번씩 반복하여 분석한 후, 그 분석 과정에서 문항 수가 30문항, 25문항, 20문항으로 변함에 따라 검사의 신뢰도는 어떻게 변하는지 또한 알아보았다.

연구 결과를 요약하면 다음과 같다. 첫째, 동등배점 조건에서 검사의 신뢰도는 차등배점 조건에서 검사의 신뢰도보다 일반적으로 높게 나타났다. 이는 차등배점을 적용할 특별한 이유와 근거가 없는 한 검사의 신뢰도 측면에서는 동등배점을 적용하는 것이 타당하다는 선행연구의 결과와 일치한다. 둘째, 차등배점 조건 내에서 배점 간 점수 차이가 커질수록 검사의 신뢰도는 감소하였다. 따라서 차등배점을 적용하여 검사 문항을 구성하더라도 배점 차이가 크지 않도록 조정하는 것이 검사의 신뢰도를 높이기 위한 측면에서 중요하다고 할 수 있다. 셋째, 문항의 수가 줄어들수록 신뢰도는 비교적 크게 감소하였다. 피험자 분포가 정규분포를 이루는 경우 문항의 수에 의해 신뢰도가 감소하더라도 적정한 수준의 신뢰도 .80을 만족하였지만, 피험자 분포가 부적편포일 경우에는 문항 수가 20문항일 때, 신뢰도가 .80이하로 감소하였다. 이는 부적편포가 있는 피험자 집단의 경우 검사의 신뢰도가 적정 수준 이상을 유지하기 위하여 적어도 25문항 이상을 만족하는 것이 필요하다는 것을 보여준다.

이 연구는 차등배점을 고려한 혼합형 검사에서 차등배점 조건과 피험

자 분포, 검사 문항 수가 검사의 신뢰도에 어떤 영향을 미치는지를 일반화가능도 이론을 적용하여 분석하였다는 점에서 의의가 있다. 추가적으로, 이 연구에서 사용한 일반화가능도 설계는 교과별 검사의 안정적인 신뢰도를 확보하기 위하여 효율적인 문항 배점 방식, 문항 수 등을 예시하는데 활용할 수 있다. 또한 종합적인 평가 점수의 신뢰도를 높이기 위한 각 하위 검사의 가중치 부여 방식 등의 분석에도 적용할 수 있을 것으로 기대된다.

이 연구의 제한점은 실제 자료가 아닌 모의실험 자료를 생성하였기 때문에 실제 자료에서는 다르게 나타날 수 있는 피험자 분포 등의 오차 요인을 고려하지 못하였다는 것이다. 또 배점 방식을 선형적으로만 변화시켰기 때문에 다양한 배점 방식에 따른 신뢰도를 분석하지 못했다는 점에서 한계를 가진다. 따라서 후속 연구로 실제 자료를 바탕으로 조금 더 다양한 오차 요인과 차등배점 부여방식의 다양성을 고려하여 분석을 진행한다면 보다 의미 있는 연구가 될 것으로 생각된다.

주요어 : 신뢰도, 차등배점, 일반화가능도 이론, 모의실험

학 번 : 2014-20841

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1. 연구의 필요성 및 목적

우리나라의 중·고등학교 현장에서 주로 사용되는 혼합형 검사에서는 문항 유형과 배점의 다양화를 통해 검사의 변별력을 높이고 있다. 이는 등급이 부여되거나 선발이 필요한 성적관리 체계에서 동점자가 생기는 것을 방지하고, 문항의 수준과 풀이 시간 등을 고려하여 검사의 타당성을 높이기 위함이다. 그러나 교육측정학적 측면에서, 검사의 구성 요소에 따라 그 검사의 신뢰도가 달라질 수 있음을 충분히 고려하지 않는다면 다양한 유형의 문항과 배점방식은 오히려 신뢰도를 떨어뜨려 검사의 타당성마저 위협할 수 있다.

서울시교육청(2015a, 2015b)은 중학교 및 고등학교 학업성적관리 시행 지침을 통해 개별학교 수준의 지필평가에서 상위 동점자 발생을 사전에 예방하기 위하여 평가 문항을 늘리고 수준별 난이도에 유의하여 배점을 다양하게 할 것을 제안하고 있다. 그러나 이와 관련하여 구체적인 가이드라인이나 교과별 특성을 고려한 문항 유형의 비율과 배점에 대한 예시는 찾기는 어렵다. 이현숙(2012)의 연구에서는 개별 학교에서 교사들이 제작하는 중간고사나 기말고사의 경우, 문항 유형별 가중치가 체계적인 연구 절차 없이 경험적으로 교과 교사들의 판단이나 과거에 실시된 시험 관례를 따르는 경향이 있음을 지적하였다.

개별학교뿐만 아니라 국가 수준에서 실시하는 여러 시험에서도 검사를 구성하는 문항 유형의 비율이나 차등배점과 관련된 연구가 충분하다

고 보기 어렵다. 김주훈 외(2010)의 연구에서 차등배점을 적용할 때의 논리적 일관성과 타당성을 분석하기 위해 국가고시의 배점 방식을 조사한 결과, 차등배점은 동점자 해소 이외의 긍정적인 효과를 기대하기 어렵고 임의성이 개입될 소지가 많다는 점이 지적되었다. 같은 맥락에서 양길석(2007)의 연구는 국가수준의 한국어능력시험 자료를 바탕으로 차등배점의 기준이 임의적일 수 있으며, 기준을 적용할 때에 경험적 자료보다는 선험적 판단에 의존하기 때문에 판단의 신뢰성을 확보하기 어려움을 주장하였다. 또한 김신영, 노국향(1999)의 연구에서는 대학수학능력시험 자료를 바탕으로 선택형 검사에서 차등배점은 특별한 이유와 근거가 없는 한 동등배점을 부여하는 것이 타당하다고 언급하였다.

측정학적으로 검사의 양호도를 판단하기 위해 신뢰도를 분석하는 방식에는 재검사신뢰도, 동형검사신뢰도, 내적일관성계수 등 고전검사이론에 따른 방법이 활발히 사용되고 있으나, 일반화가능도 이론(Generalizability theory) 역시 신뢰도 분석을 위해 사용될 수 있다. 일반화가능도 이론은 다양한 종류의 측정 오차를 고려한 모형을 이용하여 분산성분의 추정을 통해 검사에 영향을 미치는 오차원의 영향력과 검사의 신뢰도를 분석할 수 있다(Brennan, 1992; 이종성, 1997). 일반화가능도 분석에서는 기존의 고전검사이론 모형에서 확장된 모형에 분산분석(ANOVA)을 적용하는 절차를 포함하는데, 이를 통해 검사의 신뢰도뿐 아니라 조금 더 효율적인 검사 설계를 제시할 수 있다는 장점을 갖는다.

일반화가능도 이론을 적용하여 검사의 신뢰도를 분석한 연구들을 살펴보면, 교과별 시험(김명화, 2005; 이기영, 안희수, 2005; Powers & Brennan, 2009), 언어 관련 능력시험(신동일, 2001; 이영식, 신상근, 2004; Lee & Kantor, 2005; Xi & Mollaun, 2006; 양길석, 2007; 김경선, 이규민, 강승혜, 2010; 이은하, 2015; Koizumi, 2015), 글쓰기(김성숙, 1995; 노

국향, 1995; Schoonen, 2005), 체육(김도연, 허종관, 2002), 의학(Clauser, Harik, & Margolis, 2006), 평가도구(김성숙, 1993; 김정환, 이용환, 1999), 자기소개서 및 추천서(김성연, 한기순, 2013) 등 여러 영역에 걸쳐 다양한 연구가 있었다. 이는 일반화가능도 이론이 채점자, 문항 유형, 평가 영역, 피험자 집단 등 다양한 오차원을 고려하여 신뢰도 계수를 구할 수 있으며, 이에 따라 가장 효율적으로 검사를 실시하기 위한 채점자 수, 평가 영역 수 등 구체적인 정보를 제시하는 장점이 있음을 보여준다.

모의자료를 이용하여 일반화가능도 분석을 진행한 연구(Briggs & Wilson, 2007; Eggen &Veldkamp, 2012)도 있다. 이 경우 문항반응이론(item response theory)을 적용하여 생성된 모의자료를 일반화가능도 이론으로 분석하여 측정 조건에 따른 신뢰도 계수나 분산을 산출하였다.

이에 따라, 이 연구에서는 검사의 여러 구성 요소 중 차등배점 기준 또한 오차요인이 될 수 있음을 고려하여, 모의로 생성된 피험자 및 문항 자료를 바탕으로 배점 방식에 따라 검사의 신뢰도가 어떻게 달라지는지 일반화가능도 이론을 적용하여 분석하였다. 이를 통해, 정규분포를 따르는 피험자 집단을 가정했을 때와 상위 동점자 분포가 생길 수 있는 부적편포를 따르는 피험자 집단을 가정했을 때, 여러 배점 조건에 따라 피험자 및 문항의 분산 추정치와 신뢰도 계수가 어떻게 달라지는지 살펴보았다. 또한 문항의 수를 조절하였을 때의 신뢰도 계수의 변화를 분석하여 검사가 양호한 수준의 신뢰도를 만족하기 위해서는 배점 조건과 문항 수가 어느 정도가 되어야 하는지를 제안하고자 하였다.

차등배점이 검사를 구성하는 문항들의 가중치라는 점을 고려할 때, 각 문항의 점수를 합산하여 검사점수를 구하는 것과 여러 가지 다른 종류의 검사 점수를 어떻게 종합하느냐 하는 것은 근본적으로 같은 문제라 볼 수 있다(임인재, 김신영, 박현정, 2011). 따라서 이 연구의 결과는 차등배

점이 있는 혼합형 검사의 신뢰도뿐만 아니라 여러 검사 점수의 가중치를 고려한 종합적인 평가의 신뢰도를 구하는 방식에도 응용될 수 있을 것으로 기대된다.

2. 연구 문제

이 연구에서는 중·고등학교 현장에 적용되는 혼합형 검사임을 가정한 모의자료에서 배점 방식에 따라, 구체적으로는 동등배점 및 여러 차등배점의 조건에 따라 검사의 신뢰도가 어떠한 차이를 보이는지 일반화가능도 이론을 적용하여 분석하고자 하였다. 일반화가능도 이론의 G연구를 통해 여러 조건에서의 분산 추정치와 각 오차원의 비율을 확인하고, D연구를 통해 배점조건, 피험자 분포 조건, 문항 수 조건에 따라 신뢰도가 어떻게 달라지는지 보고자 하였다. 연구문제를 정리하면 다음과 같다.

연구문제 1. 혼합형 검사에서 동등배점 및 차등배점 조건을 다양하게 적용하였을 때, 조건에 따라 신뢰도는 어떠한 차이가 있는가?

연구문제 2. 혼합형 검사에서 피험자의 분포가 부적편포를 이룰 때, 동등배점 및 차등배점 조건에 따라 신뢰도는 어떠한 차이가 있는가?

연구문제 3. 혼합형 검사에서 문항 수 조건을 변화시켰을 때, 조건에 따라 신뢰도는 어떠한 차이가 있는가?

Ⅱ. 이론적 배경

1. 차등배점에 대한 논의

검사를 구성하는 여러 문항의 점수를 종합하는 방법은 크게 임상적 방법, 합리적 방법, 경험·통계적 방법으로 나뉜다. 이 중 경험·통계적 방법은 검사점수가 최대의 타당도를 갖도록 각 문항에 차별적 가중치를 주는 방법을 말한다(임인재, 김신영, 박현정, 2011). 이러한 가중치의 동일 여부에 따라 문항 배점 방식은 동등배점과 차등배점으로 구분된다(양길석, 2007; 김주훈 외, 2010).

동등배점은 모든 문항에 동일한 가중치를 주어 문항배점을 결정하는 것으로 k개의 문항으로 이루어진 검사에서 동등배점을 적용한 검사의 검사점수는 다음과 같다.

$$X = I_1 + I_2 + I_3 + \cdots + I_k$$

동등배점을 적용한 검사는 채점 절차와 해석이 비교적 용이하나, 실제적으로 문항 내용의 중요도, 문항의 난이도, 사고력 수준, 문제풀이 시간 등 문항 간 차별적인 속성을 고려할 수 없다는 단점을 지닌다(양길석, 2007).

차등배점은 동등배점과 달리 개별 문항의 점수에 가중치를 다르게 하여 검사 점수를 종합하는 것으로 k개의 문항에 대해 w의 가중치를 준 검사의 검사점수는 다음과 같다.

$$X = w_1 I_1 + w_2 I_2 + w_3 I_3 + \cdots + w_k I_k$$

차등배점은 검사를 구성하는 각 문항의 수준과 내용이 동일하지 않다고 가정한다.

각 문항의 가중치를 부여하는 방식은 여러 가지가 있으나(Russell, Hubley, & Zumbo, 2006; 김신영, 노국향, 1999) 우리나라의 학습 상황에서 사용되는 대부분의 성취도 검사는 전문가 판단방법에 의해 문항 배점이 부여된다(김신영, 노국향, 1999). 전문가 판단방법은 전문가가 문항의 난이도와 중요도를 선험적으로 판단하여 배점을 부여하는 방식으로, 전문가가 피험자와 문항의 특성을 충실히 파악하여 판단할 경우 검사의 활용 목적에 적합한 검사 결과를 제공하고, 다른 방법이 다룰 수 없는 독특한 형식의 변별을 가능하게 한다는 장점이 있다. 그러나 판단의 준거가 명료하지 않을 경우 객관성이 약하고 임의적 측정이라는 비판을 면하기 어렵다(양길석, 2007).

차등배점을 적용한 검사는 피험자가 받을 수 있는 원점수의 경우의 수를 증가시켜 보다 세밀하게 피험자를 변별할 수 있다. 그러나 차등배점의 타당성과 관련된 선행연구를 살펴보면 피험자를 변별하여 동점자 문제를 해결할 수 있는 것 이외의 긍정적 효과는 그다지 보고되지 않았다. 김신영, 노국향(1999)의 연구에서는 대학수학능력시험 자료를 활용하여 선택형 문항에 차등배점을 부여할 경우 적용되는 차등배점의 논리와 타당성에 대해 분석하였다. 그 결과 교과 전문가의 상호간 평정 일치도는 낮았으며 동등배점을 적용한 경우와 차등배점을 적용했을 때의 검사 점수의 상관은 1에 가까운 것으로 나타났다. 이는 선택형 검사에서 차등배점은 특별한 이유와 근거가 없는 한 동등배점을 부여하는 것이 타당하

다는 것을 의미하며, 따라서 연구자들은 차등배점을 부여해야 할 경우 이에 대한 명확하고 엄격한 기준이 필요하다고 제언하였다. 이는 Russell 외(2006)의 연구에서 차등배점과 동등배점을 적용한 각각의 검사점수에서 상관이 1에 가까우며, 차등배점을 적용한 검사가 동등배점을 적용한 검사보다 더 낫다고 보기 어렵다는 분석 결과와 일치한다.

양길석(2007) 또한 차등배점을 적용한 검사는 일정한 검사 양호도를 확보하면서 점수의 가짓수를 늘린다는 장점이 있으나, 수험자를 줄 세울 필요가 없는 자격부여시험에서는 굳이 차등배점을 적용할 필요가 없음을 시사하였다. 김신영, 노국향(1999)의 연구에서도 마찬가지로 고부담 평가에서 차등배점에 따라 피험자의 등위가 바뀔 수 있다는 점을 고려할 때, 차등배점의 근거를 명확히 설정해야 할 필요가 있음을 주장하였다.

김주훈 외(2010)의 연구에서는 각종 국가고사의 배점 방식을 조사하고, 2009학년도 의학교육입문검사를 활용하여 전문가 판단 방법에 따른 차등배점의 논리와 타당성을 경험적으로 분석하였다. 분석을 통해 차등배점의 요인을 예상 정답률, 문항 풀이 시간, (교육 과정상의) 중요도의 3가지로 추출하였다. 또한 차등배점을 합리적으로 적용하기 위한 방안으로 차등배점의 장단점에 대한 명확한 이해와 전문가들을 대상으로 한 사전 교육 및 연수가 선행되어야 하며, 차등배점을 적용할 경우 최소 배점과 최대 배점의 차이를 3배 미만으로 설정해야 시험 관련 이해 당사자들의 논란을 최소화 할 수 있음을 주장하였다.

이 외에 문항 배점과 관련한 연구로는 문항 총점과 검사 기준과의 관계가 최대가 되도록 설정해야 문항 가중치를 설정해야 한다는 Bayuk(1973)의 연구, 검사 문항의 배점에 따라 제공하는 정보의 차이를 분석한 노국향, 박정(2001)의 연구, 피험자에 따라 적용되어야 하는 문항 가중치가 다를 수 있음을 지적한 Russell 외(2006)의 연구, 전문가의 판

단에 따른 문항 가중치를 문항반응이론에 적용하는 방법에 대해 논의한 박찬호, 강태훈(2011)의 연구가 있었다.

2. 고전검사이론에서의 신뢰도

고전검사이론에서 신뢰도(reliability)란 주어진 검사가 얼마나 일관적으로 피험자의 능력을 측정하고 있는지를 보여주는 지표이다. 고전검사이론에서는 주어진 검사에서 어떤 피험자의 관찰점수(observed test score)가 진점수(true score)와 오차요인(random error component)으로 구성된다고 가정하는데, 이를 식으로 나타내면 다음과 같다.

$$X = T + E$$

X는 피험자의 관찰점수를, T와 E는 각각 피험자의 진점수와 오차요인을 나타낸다. 이때. 피험자의 진점수와 오차요인의 상관은 없다고 가정한다.

신뢰도의 뜻을 생각해볼 때 피험자의 관찰점수가 진점수를 잘 반영할수록 검사의 신뢰도는 높아질 것이다. 또한 신뢰도가 높다면 같은 영역을 측정하고 문항의 난이도가 같은 두 동형검사(parallel tests)를 실시했을 때, 두 검사의 점수는 비슷하게 나타날 것이다. 즉, 고전검사이론에서 신뢰도는 관찰점수 분산에 대한 진점수 분산의 비율 또는 두 동형검사 사이의 관찰점수 상관으로 정의할 수 있다. 이에 따라 고전검사이론에서 신뢰도 계수를 산출하는 방법은 유도해보면 다음과 같다.

상관계수의 정의에 따라 총 N명의 피험자 중 i번째 피험자의 관찰점

수와 진점수의 상관은 다음과 같이 나타낼 수 있다. 이때 x_i는 i번째 피험자의 관찰점수와 전체 피험자의 관찰점수 평균의 차이, t_i는 i번째 피험자의 진점수와 전체 피험자의 진점수 평균의 차이, 즉, 편차 점수를 뜻한다.

$$\rho_{XT} = \frac{\sum_i (X_i - \overline{X})(T_i - \overline{T})}{N\sigma_X\sigma_T} = \frac{\sum_i x_i t_i}{N\sigma_X\sigma_T}$$

관찰점수의 편차점수는 진점수와 오차요인의 편차점수로 나타낼 수 있으므로, 위의 식을 다시 다음과 같이 나타낼 수 있다.

$$\rho_{XT} = \frac{\sum_i (t_i + e_i)t_i}{N\sigma_X\sigma_T} = \frac{\sum_i t_i^2}{N\sigma_X\sigma_T} + \frac{\sum_i t_i e_i}{N\sigma_X\sigma_T}$$

이때, 고전검사이론의 가정에 의해 피험자의 진점수와 오차성분의 상관이 0이므로 진점수와 관찰점수의 상관은 다음과 같이 관찰점수 표준편차에 대한 진점수 표준편차로 나타내어진다.

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X\sigma_T} = \frac{\sigma_T}{\sigma_X}$$

즉, 이와 같이 고전검사이론에서 신뢰도 계수(reliability coefficient)는 진점수와 관찰점수 상관의 제곱으로 볼 수 있다.

신뢰도 계수를 두 동형검사 간 관찰점수의 상관이라는 관점에서 유도

해볼 수도 있다. 동형검사는 그 정의에 의해 임의의 피험자가 두 검사에서 같은 진점수와 오차요인 분산을 가지고 있음을 가정한다. 즉, 두 개의 동형검사 X_1과 X_2에서 피험자의 진점수와 오차요인 분산은 다음과 같은 특성을 지닌다.

$$t_1 = t_2,\ \sigma_{X_1}^2 = \sigma_{X_2}^2$$

이때, 두 동형검사 간 관찰점수의 상관은 다음과 같다. 위의 유도과정에서와 마찬가지로 소문자료 표기된 x_1과 x_2는 피험자가 각 검사 X_1, X_2에서 얻은 관찰점수의 편차를 뜻한다.

$$\rho_{X_1X_2} = \frac{\sum(X_1 - \overline{X_1})(X_2 - \overline{X_2})}{N\sigma_{X_1}\sigma_{X_2}} = \frac{\sum x_1 x_2}{N\sigma_{X_1}\sigma_{X_2}}$$

$x_1 = t_1 + e_1,\ x_2 = t_2 + e_2$이므로, 위의 식을 전개하면 다음과 같다.

$$\rho_{X_1X_2} = \frac{\sum(t_1 + e_1)(t_2 + e_2)}{N\sigma_{X_1}\sigma_{X_2}} = \frac{\sum t_1 t_2}{N\sigma_{X_1}\sigma_{X_2}} + \frac{\sum t_1 e_2}{N\sigma_{X_1}\sigma_{X_2}} + \frac{\sum e_1 t_2}{N\sigma_{X_1}\sigma_{X_2}} + \frac{\sum e_1 e_2}{N\sigma_{X_1}\sigma_{X_2}}$$

고전검사이론의 가정과 동형검사의 가정에 따라 피험자의 진점수와 오차요인과의 상관은 0이고, 두 동형검사의 각 오차요인 간의 상관 또한 0이다. 두 검사에서 진점수와 오차요인 분산은 같으므로($t_1 = t_2,\ \sigma_{X_1}^2 = \sigma_{X_2}^2$), 두 검사 간의 상관은 다음과 같이 정리된다.

$$\rho_{X_1X_2} = \frac{\sum t_1 t_2}{N\sigma_{X_1}\sigma_{X_2}} = \frac{\sum t_1^2}{N\sigma_{X_1}^2} = \frac{\sigma_T^2}{\sigma_X^2}$$

즉, 이와 같이 고전검사이론에서 신뢰도 계수는 두 동형검사 간 상관이며 관찰점수 분산에 대한 진점수 분산의 비율이라고 할 수 있다.

3. 일반화가능도 이론

1) 일반화가능도 이론에서의 신뢰도

고전검사이론에서 오차요인을 피험자의 관찰점수에서 진점수를 뺀 값으로 나타냈던 것과 달리, 일반화가능도 이론에서는 상황에 따라 다양한 오차요인을 반영하여 피험자의 능력을 추정할 수 있다. 일반화가능도 이론에서는 기존의 측정 모형인 고전검사이론에서의 모형에서 오차요인을 다양하게 구분하여 분산분석(ANOVA)을 적용한다. 그러나 분산분석에서 특정 효과의 유의한 정도를 분석했던 것과 달리 일반화가능도 이론은 다양하게 고려한 오차요인의 분산을 계산하여 상대적인 영향력을 파악하고 신뢰도 계수를 제공한다. 따라서 효율적인 검사를 위하여 여러 측정 조건에 대한 유용한 정보를 제공한다는 점이 큰 장점으로 인식된다 (Brennan, 1992; 김성숙, 1995).

구체적으로, 일반화가능도 분석은 일반화가능도(Generalizability) 연구인 G연구와 결정(Decision) 연구인 D연구로 나눠진다. G연구의 목적은 허용가능한 관찰전집(universe of admissible observations)에 대해 분산성분을 추정하는 것으로 이를 통해 여러 오차요인의 상대적인 영향력을

파악할 수 있다. D연구는 G연구에서 산출한 분산 성분을 바탕으로 어떤 국면(facet)에 따라 해당 검사의 상대오차와 절대오차, 일반화가능도 계수와 의존도 계수를 제공하여 연구자가 측정 절차에 대해 의사결정을 내릴 수 있도록 돕는다(Brennan, 2001).

일반화가능도 이론에서 사용되는 용어를 살펴보면, 국면이란 일반화가능도 분석을 하고자 할 때 문항 수나 채점자 수와 같이 연구자가 관심을 갖는 특정한 조건을 뜻한다. 허용가능한 관찰전집이란 측정대상인 모집단(population)과 한 개 이상의 국면인 전집(universe)을 포함한 것으로 G연구에서 일반화하고자 하는 조건의 범위를 뜻한다. 일반화 전집(universe of generalization)은 허용가능한 관찰전집의 부분집합으로, D연구에서 연구자가 일반화하고자 하는 전집이며 허용가능한 관찰전집에 포함된다.

일반화가능도 이론의 장점은 고전검사이론에서는 실제로 제작하기 어려운 동형검사가 존재하고 오차요인 간의 상관이 0이라는 엄격한 가정을 적용하는 것과 달리, 측정 점수의 분포 형태나 전제 조건에 대하여 강한 가정을 필요로 하지 않는다(김성숙, 1992; 이종성, 1997; 김도연, 허종관, 2002)는 것이다. 또한 사전에 계획하여 수집한 자료가 아니더라도 연구설계로의 적용이 가능하며, 분포 형태에 관한 적절한 가정이 있을 경우 보다 강력한 결과를 얻을 수 있는 장점을 가진다(김성숙, 1992; 이종성, 1997). 또 다른 장점으로 양지승과 이규민(2007)은 단위검사로 구성된 검사 점수의 신뢰도를 추정하는 경우, 일반화가능도 이론이 고전검사이론의 신뢰도 계수인 Cronbach's α가 신뢰도를 과대 추정하는 문제를 완화시킬 수 있다는 점을 보였다.

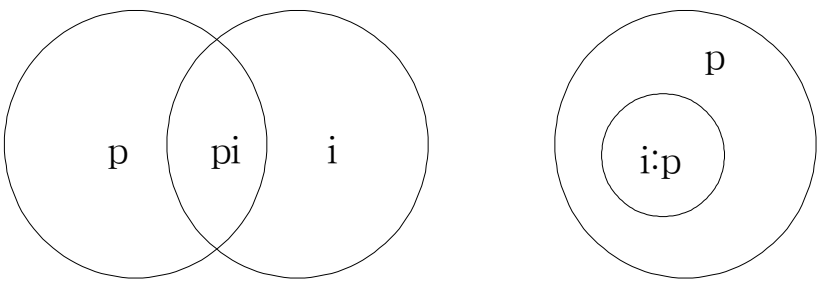
일반화가능도 분석은 연구자가 분석에 포함하고자 하는 국면과, 측정 국면들이 교차되어 있는지, 혹은 내재되어 있는지의 설계에 따라 연구자

의 관심에 맞는 여러 가지 모형을 설정할 수 있다. 이 연구에서 적용하고자 한 모형은 단변량 단일국면설계인 $p\times i$설계와 다변량 설계인 $p^{\bullet}\times i^{\circ}$설계이며, 이를 중심으로 일반화가능도 이론에 대하여 살펴보고자 한다.

2) 일반화가능도 이론의 단일국면설계

단일국면설계(single-faceted universe)는 일반화가능도 이론에서 피험자의 점수에 영향을 미치는 국면을 한 가지로 설정한 모형으로, 그 국면을 i라 할 때, 교차설계인 $p\times i$ 설계와 내재설계인 $i:p$ 설계가 있다. i를 문항 국면이라 정의할 때, $p\times i$설계는 모든 피험자가 모든 문항에 응답하는 자료의 설계를, $i:p$설계는 피험자에 따라 서로 다른 종류의 문항에 응답하는 설계를 의미한다. 이를 벤다이어그램으로 나타내면 [그림 Ⅱ-1]과 같다.

(a) $p\times i$ 설계 (b) $i:p$ 설계



[그림 Ⅱ-1] 단일국면설계의 벤다이어그램

각 설계에 따라 전집점수(universe score)를 구성하는 성분도 다르게 분해된다. 그림에서 (a) $p\times i$ 설계의 경우 전집점수는 다음과 같이 전체평균(μ, grand mean), 피험자 효과($\nu_p=\mu_p-\mu$, person effect), 문항 효과($\nu_i=\mu_i-\mu$, item effect)와 잔차($\nu_{p,i}=X_{p,i}-\mu_p-\mu_i+\mu$, residual effect)의 합으로 분해된다.

$$X_{p,i}=(\mu)+(\mu_p-\mu)+(\mu_i-\mu)+(X_{p,i}-\mu_p-\mu_i+\mu)$$

(b) $i:p$ 설계의 경우 전집점수는 전체평균(μ, grand mean), 피험자 효과($\nu_p=\mu_p-\mu$, person effect)와 잔차($\nu_{p,i}=X_{p,i}-\mu_p$,, residual effect)의 합으로 분해된다.

$$X_{p,i}=(\mu)+(\mu_p-\mu)+(X_{p,i}-\mu_p)$$

이때, 각 효과의 피험자와 문항에 대한 평균은 모두 0으로 가정한다.

$$E_p(\nu_p)=E_i(\nu_i)=E_p(\nu_{p,i})=E_i(\nu_{p,i})=0$$

모형에 따라 각각의 성분에 대한 분산(variance component) 또한 다음과 같이 분해될 수 있다.

(a) $p\times i$ 설계:

$$\sigma^2(X_{p,i})=E_pE_i(X_{p,i}-\mu)^2=\sigma^2(p)+\sigma^2(i)+\sigma^2(p,i)$$

(b) $i:p$ 설계:

$$\sigma^2(X_{p,i}) = E_p E_i (X_{p,i} - \mu)^2 = \sigma^2(p) + \sigma^2(i:p)$$

그러나 이러한 분산 성분은 알려지지 않은 무선효과(unknown random effect)에 대한 이론적인 분산성분으로 실제 자료를 바탕으로 분산 추정치를 구할 필요가 있다.

분산성분 추정치를 구하는 과정을 $p \times i$ 설계를 중심으로 설명하면 다음과 같다. 이론적인 분산성분과 달리 실제 자료의 분산성분을 추정하는 경우 다른 통계적 표기와 마찬가지로 그리스 문자가 아닌 영어 대문자로 표기한다. 전집점수를 다음과 같은 식으로 나타낼 때,

$$X_{p,i} - \overline{X} = (\overline{X}_p - \overline{X}) + (\overline{X}_i - X) + (X_{p,i} - \overline{X}_p - \overline{X}_i + \overline{X})$$

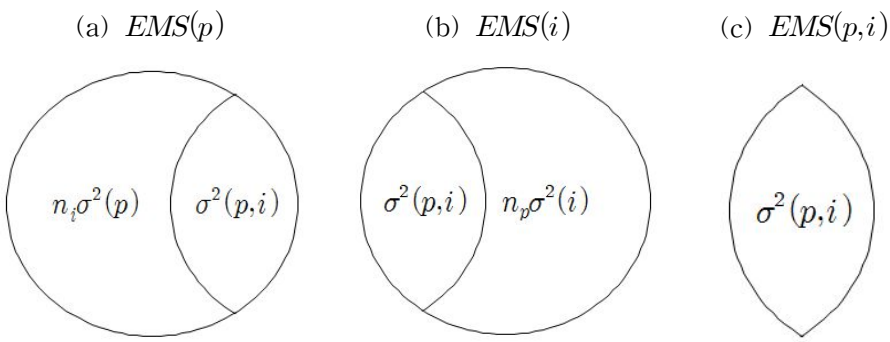
이에 대한 제곱합(sum of squares)은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} &\sum_p \sum_i (X_{p,i} - \overline{X})^2 \\ &= n_i \sum_p (\overline{X}_p - \overline{X})^2 + n_p \sum_i (\overline{X}_i - X)^2 + \sum_p \sum_i (X_{p,i} - \overline{X}_p - \overline{X}_i + \overline{X})^2 \\ &= SS(p) + SS(i) + SS(p,i) \end{aligned}$$

평균제곱합(mean squares)은 각 제곱합을 자유도(degrees of freedom)으로 나누어 준 값으로, 예를 들면 다음과 같이 표현한다.

$$MS(p) = \frac{SS(p)}{df(p)} = \frac{n_i \sum_p (\overline{X}_p - \overline{X})^2}{n_p - 1}$$

이를 바탕으로 평균제곱기댓값(expected value of mean square, EMS)을 다음과 같은 식으로 나타낼 수 있으며, 이를 직관적으로 그림으로 표현하면 [그림 Ⅱ-2]와 같다.



[그림 Ⅱ-2] 단일국면설계의 평균제곱기댓값(EMS)

$$EMS(p) = \sigma^2(p,i) + n_i\sigma^2(p)$$

$$EMS(i) = \sigma^2(p,i) + n_p\sigma^2(i)$$

$$EMS(p,i) = \sigma^2(p,i)$$

이러한 평균제곱기댓값을 통해 식을 정리하면 다음과 같은 방법으로 분산성분을 추정할 수 있다.

$$\hat{\sigma}^2(p) = \frac{MS(p) - MS(p,i)}{n_i}$$

$$\hat{\sigma}^2(i) = \frac{MS(i) - MS(p,i)}{n_p}$$

$$\hat{\sigma}^2(p,i) = MS(p,i)$$

위의 내용을 정리하여 분산성분 추정치를 구하는 공식은 정리하면 <표 Ⅱ-1>과 같다.

<표 Ⅱ-1> 단일국면설계에서 G연구의 분산성분추정치

성분	df	SS	MS	분산성분추정치
피험자 (p)	$n_p - 1$	$SS(p)$	$MS(p)$	$\hat{\sigma}^2(p) = \frac{MS(p) - MS(p,i)}{n_i}$
문항 (i)	$n_i - 1$	$SS(i)$	$MS(i)$	$\hat{\sigma}^2(i) = \frac{MS(i) - MS(p,i)}{n_p}$
상호작용 (p,i)	$(n_p - 1)(n_i - 1)$	$SS(p,i)$	$MS(p,i)$	$\hat{\sigma}^2(p,i) = MS(p,i)$

이러한 방식으로 허용가능한 관찰전집에서 분산성분을 추정하여 국면의 상대적인 영향력을 파악하고자 하는 것이 G연구의 목적이었다면, D연구는 국면의 조건에 따라 피험자의 전집점수를 일반화할 수 있도록 연구자가 어떤 결정(decision)을 내리는 분석 방법을 말한다(Brennan, 2001).

위의 단일국면설계에서의 $p \times i$ 교차설계에 대하여 D연구를 수행하는 과정을 살펴보면 다음과 같다. D연구에서는 피험자를 제외한 국면을 대

문자로 표시한다. 따라서 G연구의 $p \times i$ 설계에 이어지는 D연구의 $p \times I$ 설계에서 피험자의 전집점수 X_{pI}는 다음과 같이 분해된다.

$$X_{pI} = \overline{X_p} = \mu + \nu_p + \nu_I + \nu_{p,I}$$

이에 따라, 분산성분을 추정하는 과정은 G연구에서의 추정과정과 비슷하다. 전집점수 분산은 다음과 같이 분해된다.

$$\sigma^2(X_{p,I}) = E_p E_I (X_{p,I} - \mu)^2 = \sigma^2(p) + \sigma^2(I) + \sigma^2(p,I)$$

이때, $\sigma^2(I)$와 $\sigma^2(p,I)$는 문항의 표본 크기 또는 연구자가 설정하고자 하는 문항 수 n_i'로 나눈 값과 같다. 따라서 G연구 분산성분과 비교하여 다음과 같은 성질을 갖게 된다.

$$\sigma^2(I) = \frac{\sigma^2(i)}{n_i'}$$

$$\sigma^2(p,I) = \frac{\sigma^2(p,i)}{n_i'}$$

즉, 단일국면설계의 D연구에서 분산성분 추정치를 G연구와 비교하여 정리하면 <표 Ⅱ-2>와 같다.

<표 Ⅱ-2> 단일국면설계에서 D연구의 분산성분추정치

성분	G연구의 분산성분 추정치	D연구의 분산성분 추정치
피험자 (p)	$\hat{\sigma}^2(p) = \frac{MS(p) - MS(p,i)}{n_i}$	$\hat{\sigma}^2(p)$
문항 (i)	$\hat{\sigma}^2(i) = \frac{MS(i) - MS(p,i)}{n_p}$	$\hat{\sigma}^2(I) = \frac{\hat{\sigma}^2(i)}{n_i'}$
상호작용 (p,i)	$\hat{\sigma}^2(p,i) = MS(p,i)$	$\hat{\sigma}^2(p,I) = \frac{\hat{\sigma}^2(p,i)}{n_i'}$

이어서, D연구에서 계산한 분산성분 추정치를 바탕으로 신뢰도 계수를 구하기 위해 오차분산(error variance)을 구하는 과정은 다음과 같다. 오차분산은 준거참조평가(criterion-referenced evaluation)인지 규준참조평가(norm-referenced evaluation)인지의 평가방식에 따라 각각 절대오차분산(absolute error variance)과 상대오차분산(relative error variance)으로 나뉜다. 절대오차분산은 다음과 같이 피험자의 전집점수와 피험자 평균점수의 차이에 대한 분산으로 계산할 수 있다.

$$\Delta_{pI} \equiv X_{pI} - \mu_p = \nu_I + \nu_{pI}$$

$$\begin{aligned}\sigma^2(\Delta) &= E_I E_p(\Delta_{pI}^2) = E_I E_p((\nu_I + \nu_{pI})^2) \\ &= \sigma^2(I) + \sigma^2(pI) \\ &= \frac{\sigma^2(i)}{n_i'} + \frac{\sigma^2(p,i)}{n_i'}\end{aligned}$$

반면 상대오차분산은 학생들의 상대적인 서열에 관심이 있을 때 일반화

가능도 이론에서의 상대적인 오차에 관심을 두고자 하며 이는 다음과 같은 방식으로 계산할 수 있다.

$$\delta_{pI} \equiv (X_{pI} - E_p(X_{pI})) - (\mu_p - E_p(\mu_p)) = \nu_{pI}$$

$$\sigma^2(\delta) = E_I E_p(\delta_{pI}^2) = E_I E_p((\nu_{pI})^2)) = \sigma^2(pI)$$

$$= \frac{\sigma^2(p,i)}{n_i'}$$

즉, 절대오차분산과 상대오차분산은 다음과 같은 관계를 만족하므로,

$$\sigma^2(\Delta) = \sigma^2(\delta) + \sigma^2(I)$$

절대오차분산은 상대오차분산보다 큰 값을 가진다.

일반화가능도 이론에서 신뢰도 계수는 오차분산과 마찬가지로 규준참조평가인지 준거참조평가인지의 평가방식에 따라 각각 일반화가능도 계수(generalizability coefficient)와 의존도 계수(phi coefficient 또는 index of dependability)로 나뉜다. 이러한 계수는 전집점수 분산과 평가방식에 따른 오차분산의 합에 대한 전집점수 분산의 비로 산출할 수 있다. 즉, 고전검사이론에서 신뢰도를 산출할 때 관찰점수분산에 대한 피험자 분산의 비율을 구하는 것과 개념적으로 다르지 않다.

따라서 일반화가능도 계수는 상대오차분산을 이용하여 다음과 같이 산출한다.

$$E(\rho^2) = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$$

$$= \frac{\sigma^2(p)}{\sigma^2(p) + \frac{\sigma^2(p,i)}{n_i'}}$$

의존도 계수는 절대오차분산을 이용하여 다음과 같이 산출한다.

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}$$

$$= \frac{\sigma^2(p)}{\sigma^2(p) + \frac{\sigma_i^2}{n_i'} + \frac{\sigma_{p,i}^2}{n_i'}}$$

오차분산에서 절대오차분산이 상대오차분산보다 큰 것과 마찬가지로 일반화가능도 계수는 의존도 계수보다 크다. 또한 식을 통해 문항의 수가 많아질수록 신뢰도 계수가 증가할 것임을 예상할 수 있다.

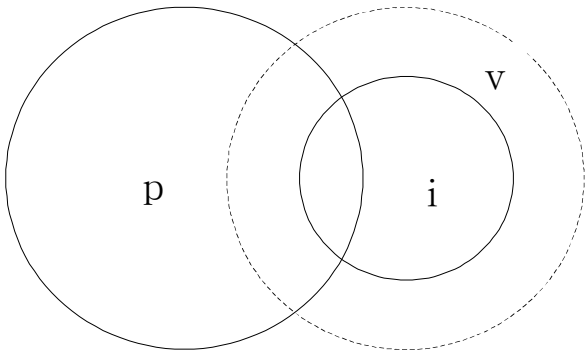
3) 다변량 일반화가능도 이론의 설계

다변량 일반화가능도 이론(multivariate generalizability theory)은 피험자가 한 개 이상의 고정된 국면에서 여러 개의 전집점수를 갖고 있고, 각 국면에 따라 무선효과 분산 성분(random-effects variance components)을 가지고 있음을 가정한다. 예를 들어, 이 연구에서 고정국면으로 활용하고자 하는 문항의 배점 조건의 국면이 3개의 수준(level)이

며, 이 수준이 변하지 않을 것이고, 이 수준에 따라 피험자가 여러 점수를 갖게 될 것임을 가정한다면 다변량 일반화가능도 분석을 적용할 수 있다.

다변량 일반화가능도 이론에서 고정국면의 설정은 단변량 일반화가능도 이론과 비교하여 분산성분의 추정을 조금 더 용이하게 한다. 이는 고정국면마다 교차되거나 내재되는 임의국면의 수가 다르더라도 단변량 일반화가능도 이론에서 불균형 설계(unbalanced design)였을 모형이 다변량 일반화가능도 이론에서는 균형 설계(balanced design)로 해석되기 때문이다. 어떤 국면의 조건 별로 문항 수가 다른 불균형 설계의 경우, 일반화가능도 이론의 분석을 실시하면 매우 복잡하다. 반면 다변량 일반화가능도 이론에서는 고정국면에 따라 분산성분을 따로 추정할 수 있고, 조건 간 공분산도 추정할 수 있다는 장점이 있다.

이 연구에서 적용한 다변량 일반화가능도 설계인 $p^{\bullet} \times i^{\circ}$ 설계를 중심으로 분산성분을 추정하는 과정을 살펴보면 다음과 같다. 우선 $p^{\bullet} \times i^{\circ}$에 쓰인 닫힌 원(•)과 열린 원(◦)은 각각 국면이 고정국면에 교차되는 것과 내재되는 것을 뜻한다. 즉, 문항의 배점 조건을 고정국면(v)으로 설정했을 때, 피험자(p)는 모든 배점 조건의 문항들을 접하게 되므로 교차되고, 문항(i)은 한 종류의 배점만을 갖게 되므로 고정국면(v)에 내재된다. 이는 단변량 일반화가능도 이론에서 $p \times (i:v)$ 설계에 대응된다고 볼 수 있다. $p^{\bullet} \times i^{\circ}$ 설계의 벤다이어그램은 [그림 Ⅱ-3]과 같다.



[그림 Ⅱ-3] 다변량 일반화가능도 설계의 벤다이어그램

$p^{\bullet} \times i^{\circ}$ 설계의 자료구조 예시는 다음 [그림 Ⅱ-4]와 같다(Brennan, 2001). 예를 들어, [그림 Ⅱ-4]와 같이 두 개의 고정국면 v_1과 v_2에 각각 6개 문항과 9개 문항이 있다면 p번째 피험자가 k번째 국면에 해당하는 i번째 문항에서 얻은 점수를 X_{p,i,v_k}라 표현할 수 있을 것이다. 이때 고정국면 v는 문항이 내재될 수 있는 국면, 예를 들어, 문항의 영역이나 문항의 유형(선다형, 서답형) 등을 표현하는 국면을 뜻한다.

	고정국면 v_1			고정국면 v_2		
p	i_1	$\cdots$	i_6	i_7	$\cdots$	i_{15}
1	$X_{1,1,v_1}$	$\cdots$	$X_{1,6,v_1}$	$X_{1,7,v_2}$	$\cdots$	$X_{1,15,v_2}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n_p	$X_{n_p,1,v_1}$	$\cdots$	$X_{n_p,6,v_1}$	$X_{n_p,7,v_2}$	$\cdots$	$X_{n_p,15,v_2}$

[그림 Ⅱ-4] 다변량 일반화가능도 설계의 자료 구조 예시

이때, 고정국면 v_1과 v_2에 대하여 피험자의 전집점수는 다음과 같이 나

타낼 수 있다. 두 식에서 ν와 ξ는 각각 고정된 v_1과 v_2국면에서 피험자의 편차점수를 뜻한다.

$$X_{p,i,v_1} = \mu_{v_1} + \nu_p + \nu_i + \nu_{p,i}$$

$$X_{p,i,v_2} = \mu_{v_2} + \xi_p + \xi_i + \xi_{p,i}$$

이와 같이, 다변량 일반화가능도 이론에서는 고정된 v_1과 v_2국면별로 피험자의 전집점수를 산출하기 때문에, 각 국면 안에서 피험자의 점수는 $p \times i$ 단일국면설계와 같으며 균형설계라 볼 수 있다. 반면, 같은 자료를 단변량 일반화가능도 설계인 $p \times (i:v)$ 모형을 이용하여 분석한다면 v국면에 따라 문항 수가 다르므로 불균형 설계가 되어 분산을 추정하는 과정이 복잡해진다. 이러한 예에서 볼 수 있듯, 다변량 일반화가능도 이론을 이용한다면 단변량 일반화가능도 설계는 다변량 일반화가능도 설계의 특수한 경우라고 볼 수 있다.

다변량 일반화가능도 설계에서 분산성분을 추정할 때는 각각의 고정국면에 해당하는 분산 성분을 추정하므로 분산-공분산 행렬을 이용한다. 즉, $p^{\bullet} \times i^{\circ}$ 설계에서 고정된 국면 v의 수준을 v_1, v_2, v_3 세 개라 가정할 때 각 성분에 따른 분산-공분산 행렬은 다음과 같다. $p^{\bullet} \times i^{\circ}$ 설계에서는 문항(i)이 고정국면(v)에 내재되어 있으므로, 문항 국면을 포함한 요인에서 공분산 추정치는 존재하지 않는다.

$$\sum p = \begin{bmatrix} \sigma^2_{v_1}(p) & \sigma_{v_1 v_2}(p) & \sigma_{v_1 v_3}(p) \\ \sigma_{v_2 v_1}(p) & \sigma^2_{v_2}(p) & \sigma_{v_2 v_3}(p) \\ \sigma_{v_3 v_1}(p) & \sigma_{v_3 v_2}(p) & \sigma^2_{v_3}(p) \end{bmatrix}$$

$$\sum i = \begin{bmatrix} \sigma^2_{v_1}(i) & 0 & 0 \\ 0 & \sigma^2_{v_2}(i) & 0 \\ 0 & 0 & \sigma^2_{v_3}(i) \end{bmatrix}$$

$$\sum p,i = \begin{bmatrix} \sigma^2_{v_1}(p,i) & 0 & 0 \\ 0 & \sigma^2_{v_2}(p,i) & 0 \\ 0 & 0 & \sigma^2_{v_3}(p,i) \end{bmatrix}$$

각 행렬의 대각(diagonal)성분은 고정국면 v_1, v_2, v_3 각각에 대한 분산성분을 나타내며, 이를 추정하는 방법은 단변량 일반화가능도 설계에서의 추정방법과 동일하다. 비대각(off-diagonal)성분은 전집점수에 대한 공분산 성분을 나타내는데, 이러한 공분산 성분을 추정하는 공식은 다음과 같다.

$$S_{vv'}(p) = \frac{n_p}{n_p - 1}\left(\frac{\sum_p \overline{X}_{pv}\overline{X}_{pv'}}{n_p} - \overline{X}_v\overline{X}_{v'}\right)$$

$$\hat{\sigma}_{vv'}(p) = S_{vv'}(p)$$

이와 같이 다변량 일반화가능도 이론에서의 G연구가 끝나면 단변량 일반화가능도 이론과 비슷한 방식으로 D연구를 진행할 수 있다. 다변량 일반화가능도 이론의 D연구에서 분산성분 추정치는 G연구의 분산성분

추정치를 국면의 수준 수, 여기서는 문항 수로 나누어 계산한다. 위의 v_1, v_2, v_3 각각의 수준에서 D연구의 분산-공분산 행렬 추정치는 다음과 같이 나타낼 수 있다.

$$\widehat{\sum} p = \begin{bmatrix} \sigma^2_{v_1}(p) & \sigma_{v_1 v_2}(p) & \sigma_{v_1 v_3}(p) \\ \sigma_{v_2 v_1}(p) & \sigma^2_{v_2}(p) & \sigma_{v_2 v_3}(p) \\ \sigma_{v_3 v_1}(p) & \sigma_{v_3 v_2}(p) & \sigma^2_{v_3}(p) \end{bmatrix}$$

$$\widehat{\sum} I = \begin{bmatrix} \frac{\sigma^2_{v_1}(i)}{n_{v_1}} & 0 & 0 \\ 0 & \frac{\sigma^2_{v_2}(i)}{n_{v_2}} & 0 \\ 0 & 0 & \frac{\sigma^2_{v_3}(i)}{n_{v_3}} \end{bmatrix}$$

$$\widehat{\sum} p, I = \begin{bmatrix} \frac{\sigma^2_{v_1}(p,i)}{n_{v_1}} & 0 & 0 \\ 0 & \frac{\sigma^2_{v_2}(p,i)}{n_{v_2}} & 0 \\ 0 & 0 & \frac{\sigma^2_{v_3}(p,i)}{n_{v_3}} \end{bmatrix}$$

이를 바탕으로 각 문항 수에 따라 가중치를 고려하여 합산하면 전집점수의 분산성분 추정치와 오차성분의 분산성분 추정치를 계산할 수 있다.

전집점수의 합산된 분산성분 추정치는 다음과 같다.

$$\hat{\sigma}_C^2(p) = \sum_v w_v^2 \sigma_v^2(p) + \sum_{v \neq v'} w_v w_{v'} \sigma_{vv'}(p)$$

예를 들어, 고정국면의 수준이 3개이고 각 국면에 해당하는 문항이 각각 2개, 3개, 5개라 할 때, 전집점수분산은 다음과 같이 계산할 수 있는 것이다.

$$\hat{\sigma}_C^2(p) = \left(\frac{2}{10}\right)^2 \sigma_1^2(p) + \left(\frac{3}{10}\right)^2 \sigma_2^2(p) + \left(\frac{5}{10}\right)^2 \sigma_3^2(p)$$
$$+ 2\left(\frac{2}{10}\right)\left(\frac{3}{10}\right)\sigma_{12}(p) + 2\left(\frac{2}{10}\right)\left(\frac{5}{10}\right)\sigma_{13}(p) + 2\left(\frac{3}{10}\right)\left(\frac{5}{10}\right)\sigma_{23}(p)$$

상대오차분산과 절대오차분산 추정치 또한 비슷한 방법으로 가중치를 고려하여 계산할 수 있다.

$$\hat{\sigma}_C^2(\delta) = \sum_v w_v^2 \sigma_v^2(\delta) = \sum_v \left(\frac{w_v^2}{n_{iv}}\right) \sigma_v^2(p,i)$$

$$\hat{\sigma}_C^2(\Delta) = \sum_v w_v^2 \sigma_v^2(\Delta) = \sum_v \left(\frac{w_v^2}{n_{iv}}\right) [\sigma_v^2(i) + \sigma_v^2(p,i)]$$

일반화가능도 계수와 의존도 계수를 구하는 방식은 단변량 일반화가능도 이론에서와 같다.

$$E(\rho^2) = \frac{\sigma_c^2(p)}{\sigma_c^2(p) + \sigma_c^2(\delta)}$$

$$\Phi = \frac{\sigma_c^2(p)}{\sigma_c^2(p) + \sigma_c^2(\Delta)}$$

추가적으로, 이러한 분산의 합산 방식 때문에 각 국면마다 문항의 수가 같다면 단변량 일반화가능도 분석을 적용하는 것이 계산하기 더 쉬울 수 있다. 그러나 일반적으로 국면의 수준에 따라 문항의 수가 다를 때에는 다변량 일반화가능도 분석을 적용하는 것이 더 낫다고 알려져 있다 (Brennan, 1992).

Ⅲ. 연구 방법

1. 연구자료

이 연구에서는 혼합형 검사에서 차등배점 조건을 고려하여 신뢰도를 분석하고자 하였다. 이를 위해 혼합형 검사의 구조를 설정하고, 배점 조건을 포함하여 분석에 고려할 여러 조건들을 결정한 후, 모의자료를 생성하여 일반화가능도 분석을 적용하였다.

우선 혼합형 검사의 경우 개별 학교마다 검사의 구성이 모두 다르므로 이 연구에서는 객관성과 대표성을 확보하기 위하여 매년 국가 수준에서 실시되는 대학수학능력시험의 수리영역의 문항 구성을 바탕으로 조건을 변화시키고자 하였다. 한국교육과정평가원(2016. 3. 29.)은 2017학년도 대학수학능력시험 시행기본계획에서 수리영역의 경우 단답형 30%를 포함하여 100점을 총점으로 30문항을 출제할 것임을 밝혀두고 있다. 가형과 나형 중 피험자의 전공계열에 따라 한 개를 선택하여 응시할 수 있으며 응시 시간은 100분으로 계획되어 있다. 차등배점과 관련하여 문항별 배점은 2, 3, 4점으로 하되, 문항의 중요도와 난이도, 문항 풀이에 소요되는 시간, 사고 수준을 고려하여 배점을 부여할 것임을 또한 밝혀두고 있다.

구체적으로 지난 8년 간 대학수학능력시험 수리영역의 문항 구성을 살펴보면 <표 Ⅲ-1>과 같다. 최근 8년 간 수리영역의 문항은 가형과 나형 모두 선택형 21문항과 단답형 9문항으로, 각 문항 유형의 총점은 각각 68점과 32점이 되도록 구성되었다. 1번 문항에서 21번 문항까지 선택

형 문항이 모두 배치된 후, 22번 문항에서 30번 문항까지는 단답형 문항을 배치하였고, 문항의 배점은 문항 순서가 뒤에 있을수록 4점짜리 문항이 많았으나, 모든 문항에 해당되는 것은 아니었다. 각 배점에 따라서는 2점짜리 문항이 3문항, 3점짜리 문항이 14문항, 4점짜리 문항이 13문항으로 구성되었다.

<표 Ⅲ-1> 최근 8년간 대학수학능력시험의 수리영역 문항 구성

(문항 수, 문항 번호 예시)

구분	2점	3점	4점
선택형(21문항)	3문항 (1번～3번)	10문항 (4번～13번)	8문항 (14번～21번)
단답형(9문항)	-	4문항 (22번～25번)	5문항 (26번～30번)
전체(30문항)	3문항	14문항	13문항

피험자 규모는 100명, 300명의 2가지 조건을 바탕으로 자료를 생성하고자 하였다. 이는 중·고등학교의 혼합형 검사임을 고려하였을 때, 한 학년의 수가 적거나 일부 학급을 대상으로 성적을 산출하는 경우 100명 정도의 피험자가 시험에 응시하고, 큰 규모의 학교라 하더라도 특정 시험에 응시하는 학생 수가 300명 정도일 것임을 가정하였기 때문이다. 문항 분석과 관련하여 모의실험을 적용한 기존의 여러 선행연구(Kim & Lee, 2004; 이문수, 이규민, 강상진, 2009)에서 300명이나 500명을 최소 규모로 설정한 것에 비하면 이 연구의 피험자 수는 큰 규모가 아닐 수 있다. 그러나 일반화가능도 이론을 이용하여 모의실험을 진행한 연구가 많지 않고, 일반화가능도 이론은 피험자 수가 비교적 적은 경우에도 분

석이 진행된다는 점을 고려하면 이 연구가 분석하고자 하는 대상의 특성에 맞게 피험자 규모를 설정하였다고 볼 수 있을 것이다.

문항 배점 조건은 앞서 언급한 대학수학능력시험의 수리영역 문항 구성을 변형하여 설계하였다. 30개의 문항에 동일하게 3.3점을 부여하여 총점을 99점으로 하는 동등배점 조건과, 30개의 문항에 2, 3, 4점의 차등배점을 부여하는 대학수학능력시험의 수리영역 구성을 포함하여 3개의 차등배점 조건을 문항 배점 조건으로 설정하였다. 차등배점 조건은 배점별 문항의 수와 총점을 고정하고, 문항 난이도에 따라 배점의 차이가 최소 0.4점에서 최대 1.6점이 되도록 설계하였다. 이 연구에서 고려한 문항 배점별 조건은 <표 Ⅲ-2>와 같다.

<표 Ⅲ-2> 문항 배점별 조건

구분	저배점 (3문항)	중배점 (14문항)	고배점 (13문항)	총점 (30문항)
조건a (동등배점)	3.3점	3.3점	3.3점	99점
조건b	2.8점	3.2점	3.6점	100점
조건c (현행수능)	2.0점	3.0점	4.0점	100점
조건d	1.2점	2.8점	4.4점	100점

피험자 분포 조건은 차등배점을 적용한 검사가 상위권 피험자의 변별을 목적으로 한다는 점을 고려하여(서울시교육청, 2015a; 서울시교육청, 2015b) 피험자 분포가 정규분포를 만족할 때와 부적편포를 이룰 때 신뢰도가 어떻게 달라지는지 알아보고자 하였다. 정규성을 위배하는 경우에

대한 모의실험연구를 살펴보면, 신택수(2016)의 연구에서는 왜도 2와 첨도 7, 왜도 2.8과 첨도 13의 조건을 기준으로 분포를 생성하였고, Curran, West, & Finch(1996)의 연구에서는 왜도 2와 첨도 7, 왜도 3과 첨도 21의 조건을 기준으로 분포를 생성하였다. 이를 바탕으로 이 연구에서는 부적편포를 이루는 피험자 분포에서의 배점 조건에 따른 신뢰도의 변화를 살펴보기 위해 정규분포(왜도 0과 첨도 0)를 이루는 경우와 약간의 부적편포를 이루는 경우(왜도 -2와 첨도 7)를 기준으로 피험자 분포 자료를 생성하였다.

문항 수 조건은 문항 자료의 생성 후 일반화가능도 분석의 D연구에서 적용할 수 있는 조건으로 기존에 구성한 30개의 문항에 대한 분석결과를 포함하여, 문항 수를 25문항과 20문항으로 감소시키도록 하였을 때, 이에 따라 달라지는 신뢰도 계수를 살펴보고자 하였다. 다변량 일반화가능도 이론의 적용을 위하여 각 국면의 수준에 따라 문항 수를 다르게 설정할 필요가 있기 때문에 30문항의 경우 기존에 참고한 자료와 같이 저배점 3문항, 중배점 14문항, 고배점 13문항으로 유지하였다. 25문항의 경우 저배점 2문항, 중배점 11문항, 고배점 11문항으로, 20문항의 경우 저배점 2문항, 중배점, 9문항, 고배점 9문항으로 해당 문항의 비율을 어느 정도 유지하면서 저배점 문항이 적어도 2문항은 유지되도록 조건을 설정하였다. 문항 수 조건은 <표 Ⅲ-3>과 같다. 이때, 30문항의 경우 G연구에서 분산을 추정하고자 하는 자료이므로 <표 Ⅲ-1>의 문항 구성 내용에 따라 선택형과 단답형 문항 자료를 달리하여 생성하지만, D연구에서 문항 수를 고려하여 오차분산과 신뢰도를 산출하고자 하는 경우, 각 국면에 해당하는 선택형, 단답형 문항의 유형 별 비율은 G연구에서와 같다고 가정하므로 특별히 그 유형별 문항 수를 고려하지는 않는다.

<표 Ⅲ-3> 문항 수 조건

구분	저배점	중배점	고배점
30문항	3문항	14문항	13문항
25문항	2문항	11문항	11문항
20문항	2문항	9문항	9문항

이에 따라, 이 연구에서 설정한 조건은 <표 Ⅲ-4>에 제시한대로 피험자 규모 2가지, 배점 조건 4가지, 피험자 능력 분포 조건 2가지, 문항 수 조건 3가지로 총 48가지이다.

<표 Ⅲ-4> 모의자료 생성 조건

구분	수준 수	수준
사례수	2가지	100명, 300명
배점	4가지	동등배점 조건 1개, 차등배점 조건 3개
분포	2가지	정규분포 가정, 부적편포 가정
문항 수	3가지	30문항, 25문항, 20문항
전체	(2×4×2×3=) 48가지	

이 연구에서 생성하고자 한 자료구조는 간략하게 나타내면 [그림 Ⅲ-1]과 같다. 배점 조건(저배점, 중배점, 고배점)을 국면으로 설정하여 각 배점 조건에 3문항, 14문항, 13문항이 포함되도록 하였다. 문항의 순서는 <표 Ⅲ-1>과 같이 선택형 문항인 1~21번까지의 문항 중 1번~3번, 4번~13번, 14번~21번 문항이 각각 저배점, 중배점, 고배점으로 설정되도

록 하였고, 단답형 문항인 22~30번까지의 문항 중 22번~25번, 26번~30번 문항이 각각 중배점과 고배점 문항으로 설정되도록 하였다.

	저배점(3문항)			중배점(14문항)						고배점(13문항)					
p	M 1	M 2	M 3	M 4	…	M 13	C 22	…	C 25	M 14	…	M 21	C 26	…	C 30
1	X	X	X	X	…	X	X	…	X	X	…	X	X	…	X
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮
n_p	X	X	X	X	…	X	X	…	X	X	…	X	X	…	X

※ M은 선택형(multiple choice), C는 서답형(constructed response) 문항을 뜻함.

[그림 Ⅲ-1] 모의실험설계의 자료 구조

2. 일반화가능도 설계

위의 생성한 자료를 바탕으로 일반화가능도 이론을 적용하여 G연구에서의 분산성분을 추정한 후, D연구의 문항 조건을 달리함에 따라 오차분산 및 신뢰도 계수가 어떻게 달라지는지 알아보았다. 연구문제에 따라 동등배점 조건과 차등배점 조건의 신뢰도를 비교하기 위해, 또 차등배점 조건 내에서 배점 간 차이에 따라 달라지는 신뢰도를 비교하기 위해 기존의 G연구에서 사용했던 문항 수 30문항을 기준으로 신뢰도 계수를 살펴보았다. 또한 피험자분포가 정규분포가 아닌 부적편포를 이룰 경우 각 배점 조건에 따라 신뢰도는 어떻게 변화하는지 알아보았다. 이후, D연구를 바탕으로 문항 수가 25문항과 20문항일 때의 신뢰도 계수를 분석하였다.

동등배점의 경우 모든 피험자가 같은 배점의 30개 문항에 대해 응답한 자료를 분석하는 단변량 단일국면설계인 $p \times i$ 설계를 적용하였다. 따라서 피험자의 전집점수는 다음과 같이 나타낼 수 있다.

$$X_{p,i} = (\mu) + (\mu_p - \mu) + (\mu_i - \mu) + (X_{p,i} - \mu_p - \mu_i + \mu)$$

G연구에서 각 분산성분을 추정한 후 D연구에서 문항 수에 따라 문항 국면이 포함된 분산성분을 산출하는 방식은 다음과 같다.

(a) 30문항	(b) 25문항	(c) 20문항
$\sigma^2(I) = \frac{\sigma^2(i)}{30}$	$\sigma^2(I) = \frac{\sigma^2(i)}{25}$	$\sigma^2(I) = \frac{\sigma^2(i)}{20}$
$\sigma^2(p,I) = \frac{\sigma^2(p,i)}{30}$	$\sigma^2(p,I) = \frac{\sigma^2(p,i)}{25}$	$\sigma^2(p,I) = \frac{\sigma^2(p,i)}{20}$

각각의 문항 수 조건에 따라 오차분산과 신뢰도 계수를 구하는 방법은 이론적 배경에 제시된 바와 같다.

상대오차분산: $\sigma^2(\delta) = \frac{\sigma^2(p,i)}{n_i'}$

절대오차분산: $\sigma^2(\Delta) = \frac{\sigma^2(i)}{n_i'} + \frac{\sigma^2(p,i)}{n_i'}$

일반화가능도계수: $E(\rho^2) = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$

의존도계수: $$\Phi = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\Delta)}$$

차등배점의 경우 배점 조건에 따라 문항 수가 다르기 때문에 이를 고려하여 배점 조건을 고정국면으로 설정한 후 다변량 일반화가능도 설계인 $p^{\bullet} \times i^{\circ}$ 설계를 적용하였다. 국면의 수준은 저배점(v_1), 중배점(v_2), 고배점(v_3)의 3가지로 G연구에서 각 수준을 구성하는 문항의 수는 각각 3문항, 14문항, 13문항이다. 피험자는 각 수준에서 다음과 같은 세 종류의 전집점수를 가진다. 아래 식에서 ν, ξ, ζ는 각각 v_1, v_2, v_3국면에 대한 피험자의 편차점수를 나타낸다.

$$X_{p,i,v_1} = \mu_{v_1} + \nu_p + \nu_i + \nu_{p,i}$$

$$X_{p,i,v_2} = \mu_{v_2} + \xi_p + \xi_i + \xi_{p,i}$$

$$X_{p,i,v_3} = \mu_{v_3} + \zeta_p + \zeta_i + \zeta_{p,i}$$

각 국면에서의 분산-공분산 행렬을 구한 후, D연구에 적용할 경우 각 국면에 따라 다음과 같은 형태의 행렬을 가진다.

$$\widehat{\sum} p = \begin{bmatrix} \sigma^2_{v_1}(p) & \sigma_{v_1 v_2}(p) & \sigma_{v_1 v_3}(p) \\ \sigma_{v_2 v_1}(p) & \sigma^2_{v_2}(p) & \sigma_{v_2 v_3}(p) \\ \sigma_{v_3 v_1}(p) & \sigma_{v_3 v_2}(p) & \sigma^2_{v_3}(p) \end{bmatrix}$$

$$\widehat{\sum} I = \begin{bmatrix} \frac{\sigma^2_{v_1}(i)}{n_{v_1}} & 0 & 0 \\ 0 & \frac{\sigma^2_{v_2}(i)}{n_{v_2}} & 0 \\ 0 & 0 & \frac{\sigma^2_{v_3}(i)}{n_{v_3}} \end{bmatrix}$$

$$\widehat{\sum} p, I = \begin{bmatrix} \frac{\sigma^2_{v_1}(p,i)}{n_{v_1}} & 0 & 0 \\ 0 & \frac{\sigma^2_{v_2}(p,i)}{n_{v_2}} & 0 \\ 0 & 0 & \frac{\sigma^2_{v_3}(p,i)}{n_{v_3}} \end{bmatrix}$$

이때 n_{v_1}, n_{v_2}, n_{v_3}는 <표 Ⅲ-3>의 문항 수 조건에 따라 각각 3, 14, 13 이거나 2, 11, 11, 또는 2, 9, 9의 값을 갖는다. 이후, 통합분산을 구하는 방법을 통해 피험자의 전집점수 분산과 오차분산을 계산한 후 신뢰도를 구하는 원리는 동등배점에서의 설계와 같다.

3. 모의자료 생성 및 분석 절차

모의자료 생성하고 및 분석하는 과정은 다음과 같은 순서에 따라 이루어졌다.

1) 피험자와 문항 모수 생성
2) 배점 조건에 따라 피험자의 문항반응자료 작성
3) 일반화가능도 분석 - 분산 추정치 및 신뢰도 산출
4) 반복

이 과정에서 R-3.2.5 프로그램을 활용하였으며 분포를 생성하는 과정에서 R library로 PearsonDS, gsl, moments의 세 가지 library를 사용하였다.

1) 피험자와 문항 모수 생성

피험자 모수와 문항 모수는 문항반응이론을 바탕으로 생성하였다. 피험자의 능력 분포는 정규분포를 이루는 경우 평균이 0, 표준편차가 1이 되도록, 부적편포를 이루는 경우 왜도가 -2, 첨도가 7이 되도록 자료를 생성하였다. 문항의 곤란도는 평균이 0, 표준편차가 1인 정규분포 형태로 이루어져 있음을 가정하였다. 피험자 규모가 크지 않기 때문에 문항의 변별도 모수는 1로 고정하여 응답확률을 안정적으로 추정하고자 하였고, 문항의 유형에 따라 선택형 문항의 경우 문항추측도를 0.2로, 단답형 문항의 경우 문항추측도를 0으로 고정하였다.

피험자 모수: (정규분포의 경우) $\theta \sim N(0,1)$,

(부적편포의 경우) $\theta \sim$(왜도 -2, 첨도 7인 Pearson 분포)

문항 곤란도 모수: $b \sim N(0,1)$

문항 변별도 모수: $a = 1$

문항 추측도 모수: (선택형의 경우) $c = 0.2$, (단답형의 경우) $c = 0$

문항 곤란도는 높은 배점의 문항일수록, 뒤에 있는 문항일수록 선택형과 단답형 문항이 번갈아가며 더 어려워지도록 나열하였으며, 이에 따라 피험자가 문항에 정답할 확률은 다음과 같이 나타낼 수 있다.

선택형 문항에서 피험자의 정답 확률:

$$P_1(\theta) = 0.2 + (1-0.2)\frac{1}{1+e^{-1.7(\theta-b)}}$$

단답형 문항에서 피험자의 정답 확률:

$$P_2(\theta) = \frac{1}{1+e^{-1.7(\theta-b)}}$$

이를 생성하기 위한 R 코드는 다음과 같다.

```
## 피험자 및 문항 모수 생성

    N <- 100 #피험자 규모

    student <- rnorm(N) #피험자 능력분포가 정규분포를 이루는 경우
    student <- rpearson(N, moments=c(mean=0, variance=1, skewness=-2,
kurtosis=10)) #피험자 능력분포가 부적편포를 이루는 경우

    item <- sort(rnorm(30))
```

```
    mc1 <- 0.2+(1-0.2)/(1+exp(1.7*(item[1]-student))) #선택형
    mc2 <- 0.2+(1-0.2)/(1+exp(1.7*(item[2]-student))) #선택형
    (...)
    cr29 <- 1/(1+exp(1.7*(item[28]-student))) #단답형
    cr30 <- 1/(1+exp(1.7*(item[30]-student))) #단답형

## 정답확률
    correct_prob <- data.frame(student, mc1, mc2, mc3, ..., cr29, cr30)
```

2) 배점 조건에 따라 피험자의 문항반응자료 작성

위의 과정을 통해 생성된 피험자의 정답 확률은 정답인지 아닌지의 반응 자료로 변환하여야 피험자의 검사 점수를 분석할 수 있다. 따라서 0과 1 사이에서 무선적으로 생성한 난수($\sim Unif(0,1)$)와 피험자의 정답 확률을 비교하여 피험자의 문항반응자료를 생성하였다. 예를 들어, 피험자가 어떤 문항에 정답할 확률이 0.6일 때, 무선적으로 생성된 난수가 0.4라면 정답(1)으로, 피험자의 정답 확률이 0.4일 때, 무선적으로 생성된 난수가 0.5라면 오답(0)으로 저장하도록 문항반응자료를 작성하였다.
이를 생성하기 위한 R 코드는 다음과 같다.

```
## 피험자반응(정답 1, 오답 0)
    MC1 <- (mc1 > runif(length(student)))+0
    MC2 <- (mc2 > runif(length(student)))+0
    (...)
    CR29 <- (cr29 > runif(length(student)))+0
    CR30 <- (cr30 > runif(length(student)))+0

## 문항반응자료 생성
   answer <- data.frame(student, MC1, MC2, MC3, ..., CR29, CR30)
```

```
## 각 배점에 따라 (예를 들어, b조건)
   answer_b <- data.frame(student, MC1*2.8, MC2*2.8, MC3*2.8,
MC4*3.2, MC5*3.2, ..., CR25*3.2, MC14*3.6, MC15*3.6, ..., CR30*3.6)
```

3) 일반화가능도 분석 - 분산 추정치 및 신뢰도 산출

생성된 피험자의 문항반응자료를 바탕으로 일반화가능도 이론을 적용하여 분석하였다. 계산의 편의상 G연구의 분산 추정과 D연구의 분산 추정 및 오차분산, 신뢰도 계수의 계산을 동시에 실시하였다. 이에 대한 과정은 이론적 배경에 제시하였으며, R 코드는 다음과 같은 방식으로 작성하여 실행하였다.

```
## 동등배점(a) 조건에서의 G연구

     a_score_p_total <- sum(a_score_mean^2)*30
     a_item_mean <- c(mean(MC1), mean(MC2), ..., mean(CR30))*3.3

     a_score_i_total <- sum(a_item_mean^2)*length(student)
     a_score_pi_total <- sum((answer_a[-1])^2)
     a_score_mu_total <-
(sum(answer_a[-1])/(30*length(student)))^2*30*length(student)

     a_score_p_ms <-
(a_score_p_total-a_score_mu_total)/(length(student)-1)
     a_score_i_ms <- (a_score_i_total-a_score_mu_total)/(30-1)
     a_score_pi_ms <-
(a_score_pi_total-a_score_p_total-a_score_i_total+a_score_mu_total)/((length
(student)-1)*(30-1))

## 동등배점(a) 조건에서의 D연구

     a_score_pI_sigma <- a_score_pi_sigma/30
```

```
    a_score_I_sigma <- a_score_i_sigma/30
    analysis_a_delta <- a_score_pi_sigma/30
    analysis_a_gencoef <-
a_score_p_sigma/(a_score_p_sigma+analysis_a_delta)


## 차등배점(b) 조건에서의 G연구

    b_score_v1 <- (answer_b[2]+answer_b[3]+answer_b[4])/3
    b_score_v2 <- (answer_b[5]+answer_b[6]+...+answer_b[18])/14
    b_score_v3 <- (answer_b[19]+answer_b[20]+...+answer_b[31])/13

    b_score_p_v1_ms <-
((sum(b_score_v1^2)*3-(sum(b_score_v1)/length(student))^2*3*length(stude
nt))/(length(student)-1))
    b_score_p_v2_ms <- (위와 비슷한 방식으로)

    b_score_i_v1_ms <-
((sum(answer_b[2])^2+sum(answer_b[3])^2+sum(answer_b[4])^2)/length(stu
dent)-(sum(b_score_v1)/length(student))^2*3*length(student))/2
    b_score_i_v2_ms <- (위와 비슷한 방식으로)

    b_score_pi_v1_ms <-
(sum((answer_b[2:4])^2)-sum(b_score_v1^2)*3-b_score_i_v1_ms*2)/(2*(leng
th(student)-1))
    b_score_pi_v2_ms <- (위와 비슷한 방식으로)

# 차등배점(b) 조건에서의 D연구

    b_score_v1v2 <- b_score_v1*b_score_v2
    b_score_v1v3 <- b_score_v1*b_score_v3
    b_score_v2v3 <- b_score_v2*b_score_v3

    b_score_sigma12 <-
sum(b_score_v1v2)/length(student)-((sum(b_score_v1)/length(student))*(su
m(b_score_v2)/length(student)))
    b_score_sigma13 <- (위와 비슷한 방식으로)

    b_score_composite_universe <-
((3/30)^2*b_score_p_v1_sigma+(14/30)^2*b_score_p_v2_sigma+(13/30)^2*b_s
core_p_v3_sigma+2*(3/30)*(14/30)*b_score_sigma12+2*(3/30)*(13/30)*b_scor
e_sigma13+2*(14/30)*(13/30)*b_score_sigma23)
```

```
    b_score_relative_error <-
((3/30)^2*b_score_pi_v1_sigma/3+(14/30)^2*b_score_pi_v2_sigma/14+(13/30)^
2*b_score_pi_v3_sigma/13)

    analysis_b_gencoef <-
b_score_composite_universe/(b_score_composite_universe+b_score_relative_e
rror)

    b_score_absolute_error <-
((3/30)^2*(b_score_pi_v1_sigma+b_score_i_v1_sigma)/3+(14/30)^2*(b_score_
pi_v2_sigma+b_score_i_v2_sigma)/14+(13/30)^2*(b_score_pi_v3_sigma+b_sco
re_i_v3_sigma)/13)

    analysis_b_phi <-
b_score_composite_universe/(b_score_composite_universe+b_score_absolute_
error)


## 분산성분의 추정 결과가 0보다 작을 경우 (예)

    b_score_p_v1_sigma <- (b_score_p_v1_ms-b_score_pi_v1_ms)/3
    if(b_score_p_v1_sigma<0){
      b_score_p_v1_sigma<-0
      warning("Estimate of variance is negative")
    }
```

4) 반복

위의 일반화가능도 분석을 500번씩 반복하여 분석하는 방법은 다음과 같다.

```
## 각 조건을 500번씩 반복
for(i in 1:500){

...
```

```
(위의 모든 자료 생성 및 분석 내용)
...

## 필요한 데이터 리스트 작성 후 저장 (예)
     data <- list(student=student, answer=answer, answer_a=answer_a,
answer_b=answer_b, answer_c=answer_c, answer_d=answer_d,
answer_e=answer_e, answer_f=answer_f)

## 데이터 저장
     write.csv(paste0("data_", i, ".csv")
  }
```

Ⅳ. 연구 결과

1. 기술통계

피험자 규모와 배점 조건에 따라 생성된 모의실험 자료에서 나온 피험자의 검사점수는 다음 <표 Ⅳ-1>, <표 Ⅳ-2>와 같다. <표 Ⅳ-1>과 <표 Ⅳ-2>는 각각 피험자의 능력 분포가 정규분포를 이룰 때와 부적편포를 이룰 때를 가정했을 때의 검사점수이다. 피험자의 검사점수 평균은 동등배점 조건(a)에서 55.71~57.55로 가장 높게 나타났고, 배점 차이가 크지 않은 차등배점 조건(b)에서는 동등배점 조건과 거의 일치하는 결과를 보였지만, 차등배점의 배점 차가 가장 클 경우(d), 검사점수의 평균이 50.61~52.37로 동등배점 조건에서의 검사 점수보다 약 5점정도 낮아지는 경향을 보였다. 피험자 규모에 따른 평균 점수는 크게 다르지 않은 것으로 나타났다.

피험자의 점수에 대한 표준편차는 배점 차이가 큰 조건일수록 증가하였다. 피험자 규모가 100명인 경우의 표준편차가 300명인 경우의 표준편차보다 약 0.3~0.5점정도 높았다. 구체적으로, 피험자 능력 분포가 정규분포를 이루는 경우, 피험자 수가 100명일 때 검사점수의 표준편차는 배점 차이가 커질수록 4.26점에서 4.40점으로 증가하였고, 피험자 수가 300명일 때 3.72점에서 3.90점으로 증가하였다. 피험자 능력 분포가 부적편포를 이루는 경우, 피험자 수가 100명일 때 검사점수의 표준편차는 4.19점에서 4.39점으로, 피험자 수가 300명일 때 3.88점에서 4.22점으로 증가하였다.

<표 Ⅳ-1> 모의실험 자료에서 피험자의 검사점수

(피험자 능력분포가 정규분포를 이루는 경우)

(평균, 표준편차)

	배점 조건			
	a	b	c	d
100명	56.0549 (4.2601)	55.1983 (4.2833)	53.0976 (4.3040)	50.9510 (4.3950)
300명	55.7113 (3.7218)	54.8627 (3.8031)	52.7226 (3.8427)	50.6133 (3.8987)

※ 배점 조건은 a는 동등배점, b~d는 순차적으로 배점 간 차이가 커지는 차등배점을 적용하였으며, <표 Ⅲ-2>에서 확인할 수 있다.

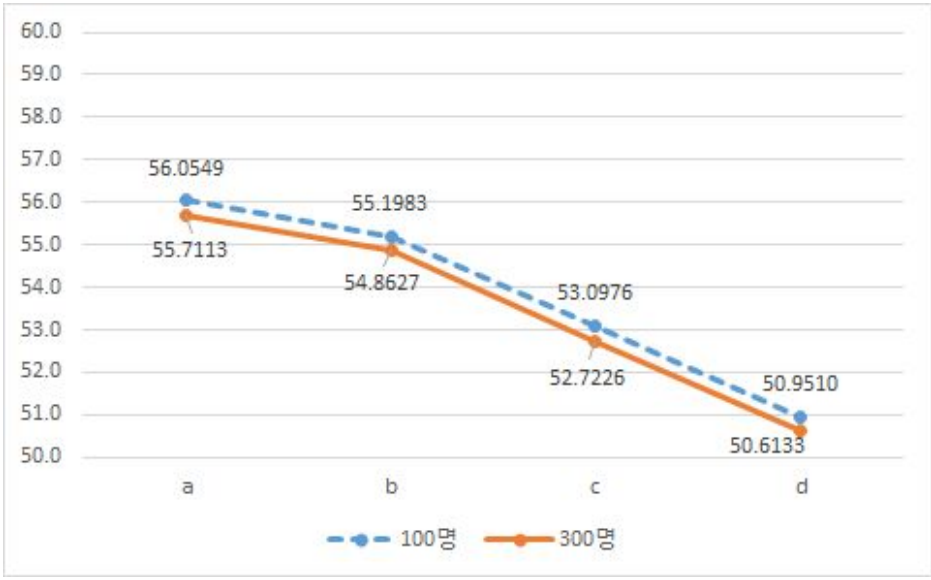
<표 Ⅳ-2> 모의실험 자료에서 피험자의 검사점수

(피험자 능력분포가 부적편포를 이루는 경우)

(평균, 표준편차)

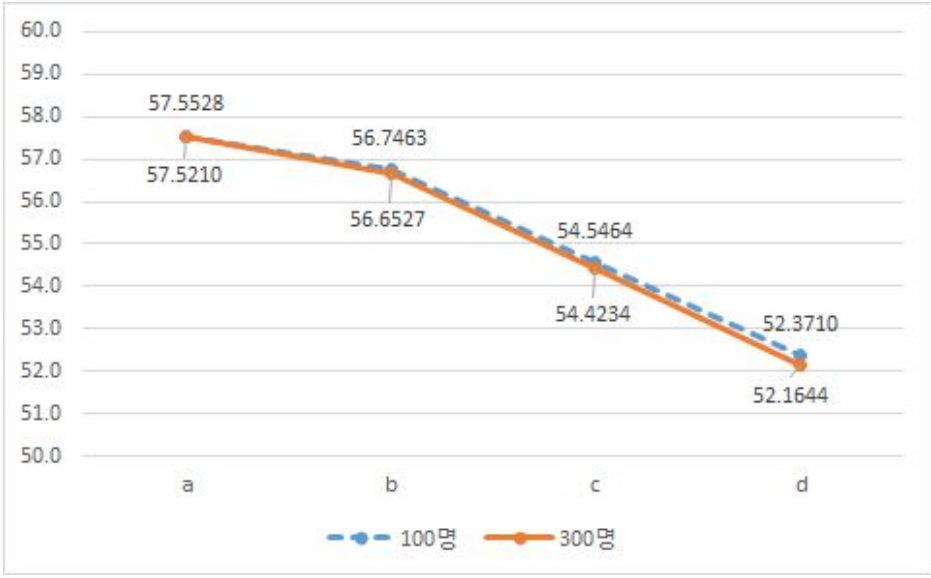
	배점 조건			
	a	b	c	d
100명	57.5528 (4.1931)	56.7463 (4.2988)	54.5464 (4.3983)	52.3710 (4.3948)
300명	57.5210 (3.8825)	56.6527 (4.0139)	54.4234 (4.1218)	52.1644 (4.2236)

※ 배점 조건은 a는 동등배점, b~d는 순차적으로 배점 간 차이가 커지는 차등배점을 적용하였으며, <표 Ⅲ-2>에서 확인할 수 있다.



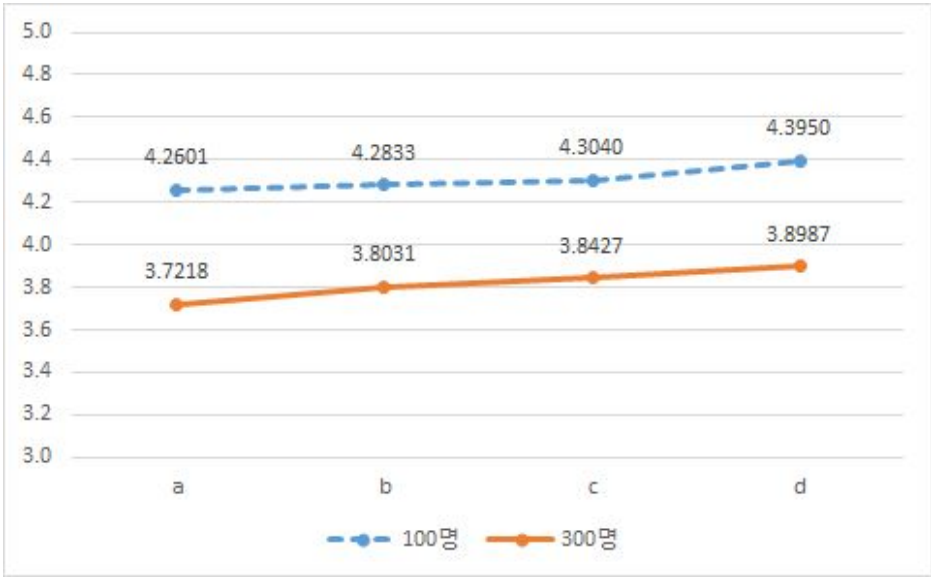
[그림 Ⅳ-1] 배점 조건에 따른 검사점수

(피험자 능력분포가 정규분포를 이루는 경우)

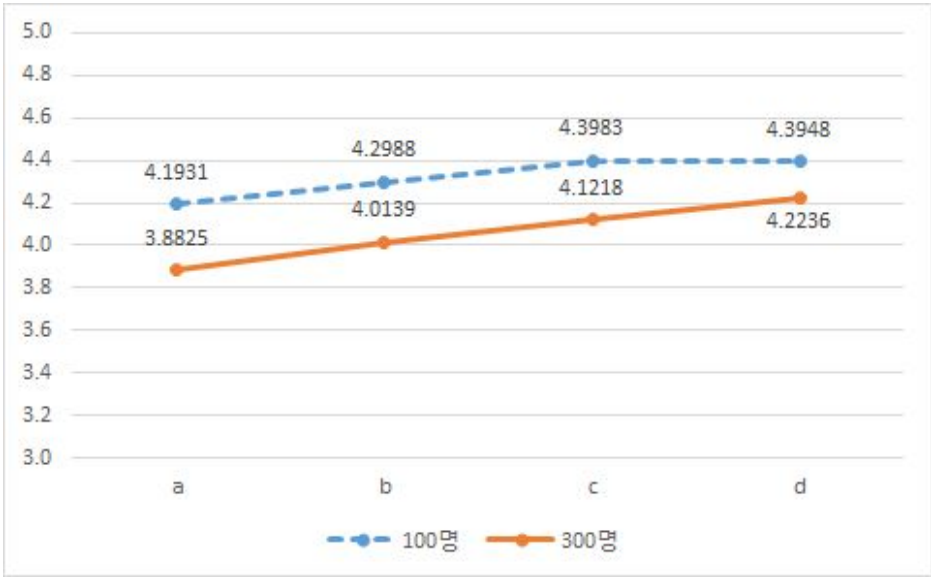


[그림 Ⅳ-2] 배점 조건에 따른 검사점수

(피험자 능력분포가 부적편포를 이루는 경우)



[그림 Ⅳ-3] 배점 조건에 따른 검사점수의 표준편차
(피험자 능력분포가 정규분포를 이루는 경우)



[그림 Ⅳ-4] 배점 조건에 따른 검사점수의 표준편차
(피험자 능력분포가 부적편포를 이루는 경우)

2. 배점 조건에 따른 신뢰도의 변화

동등배점 조건을 적용했을 때와 차등배점의 여러 조건을 적용했을 때 분산성분 추정치는 <표 Ⅳ-3>, <표 Ⅳ-4>와 같다. 동등배점 조건에서 오차요인의 영향력은 피험자에 의한 분산성분 추정치가 약 85.0~85.1%로 가장 높았고, 피험자와 문항의 상호작용에 의한 분산성분 추정치가 약 11.5~11.6%로 다음으로 높았다. 문항에 의한 분산성분 추정치는 약 3.5%로 가장 낮았다. 차등배점 조건에서 오차요인의 영향력은 피험자와 문항의 상호작용에 의한 분산성분 추정치가 약 72.8~80.0%로 가장 높았고, 피험자에 의한 분산성분 추정치가 약 17.2~23.3%로 다음으로 높았으며, 문항의 차이에 의한 분산성분 추정치는 약 1.7~9.4%로 가장 낮았다.

이를 바탕으로 30문항에 대하여 D연구의 오차분산과 신뢰도 계수를 산출한 결과를 제시하면 <표 Ⅳ-5>, <표 Ⅳ-6>과 같다. 동등배점을 적용했을 때의 일반화가능도 계수는 약 .88로 양호한 신뢰도를 나타내었다. 차등배점을 적용했을 경우 차등배점의 간격이 커질수록(b조건 → d조건) 신뢰도가 줄어들었으나 모두 약 .87이상의 신뢰도를 만족하여 피험자의 능력분포가 정규분포를 이룰 때, 일반화가능도 분석에 의한 검사의 신뢰도는 모두 양호한 수준으로 나타났다. 이를 그래프로 나타내면 [그림 Ⅳ-5]와 같다.

<표 Ⅳ-3> 100명일 때 차등배점 조건에서 G연구의 분산성분 추정치

(평균, 비율)

조건	배점	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(p,i)$
a조건	동등배점	0.4242 (84.96%)	0.0174 (3.49%)	0.0577 (11.55%)
b조건	v1 (저배점)	0.1341 (18.36%)	0.0124 (1.70%)	0.5841 (79.95%)
	v2 (중배점)	0.5150 (23.22%)	0.0847 (3.82%)	1.6179 (72.96%)
	v3 (고배점)	0.5215 (17.30%)	0.2788 (9.25%)	2.2133 (73.44%)
c조건	v1 (저배점)	0.0690 (18.34%)	0.0072 (1.92%)	0.3000 (79.74%)
	v2 (중배점)	0.4503 (23.11%)	0.0770 (3.95%)	1.4209 (72.93%)
	v3 (고배점)	0.6470 (17.37%)	0.3471 (9.32%)	2.7303 (73.31%)
d조건	v1 (저배점)	0.0253 (18.42%)	0.0025 (1.81%)	0.1095 (79.77%)
	v2 (중배점)	0.3935 (23.20%)	0.0665 (3.92%)	1.2362 (72.88%)
	v3 (고배점)	0.7762 (17.24%)	0.4184 (9.29%)	3.3078 (73.47%)

<표 Ⅳ-4> 300명일 때 차등배점 조건에서 G연구의 분산성분 추정치

(평균, 비율)

조건	배점	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(p,i)$
a조건	동등배점	0.4285 (85.08%)	0.0174 (3.46%)	0.0577 (11.46%)
b조건	v1 (저배점)	0.1371 (18.33%)	0.0139 (1.86%)	0.5971 (79.81%)
	v2 (중배점)	0.5207 (23.30%)	0.0872 (3.90%)	1.6269 (72.80%)
	v3 (고배점)	0.5222 (17.36%)	0.2813 (9.35%)	2.2044 (73.29%)
c조건	v1 (저배점)	0.0697 (18.27%)	0.0070 (1.83%)	0.3047 (79.89%)
	v2 (중배점)	0.4573 (23.28%)	0.0772 (3.93%)	1.4297 (72.79%)
	v3 (고배점)	0.6468 (17.43%)	0.3459 (9.32%)	2.7185 (73.25%)
d조건	v1 (저배점)	0.0250 (18.18%)	0.0025 (1.83%)	0.1099 (80.00%)
	v2 (중배점)	0.3990 (23.32%)	0.0658 (3.85%)	1.2461 (72.83%)
	v3 (고배점)	0.7823 (17.41%)	0.4219 (9.39%)	3.2884 (73.20%)

<표 Ⅳ-5> 100명일 때 차등배점 조건에서 D연구의 오차분산 및 신뢰도 계수

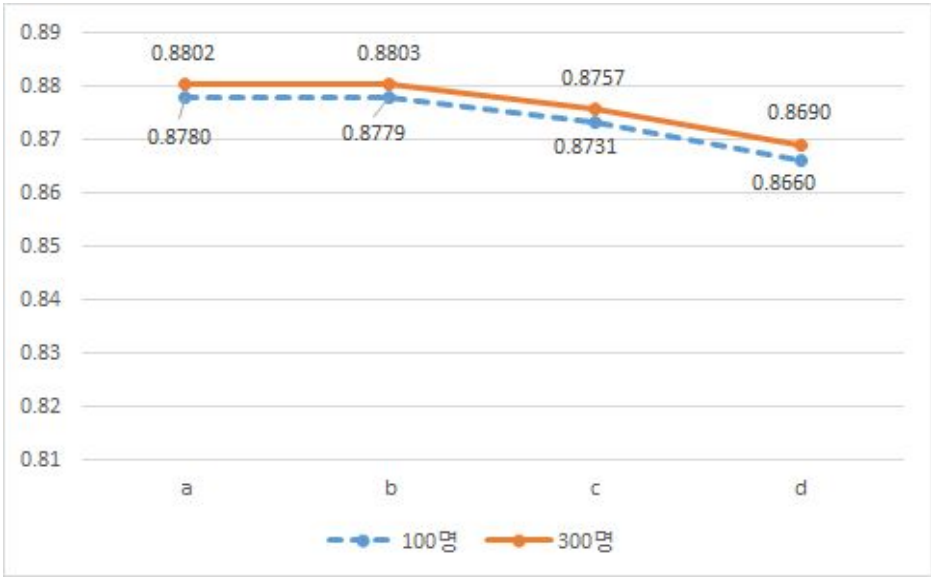
(평균, 표준편차)

배점조건	상대오차분산	일반화가능도 계수
a조건	0.0577 (0.0029)	0.8780 (0.0180)
b조건	0.0591 (0.0028)	0.8779 (0.0189)
c조건	0.0625 (0.0026)	0.8731 (0.0196)
d조건	0.0674 (0.0026)	0.8660 (0.0208)

<표 Ⅳ-6> 300명일 때 차등배점 조건에서 D연구의 오차분산 및 신뢰도 계수

(평균, 표준편차)

배점조건	상대오차분산	일반화가능도 계수
a조건	0.0577 (0.0023)	0.8802 (0.0116)
b조건	0.0591 (0.0022)	0.8803 (0.0117)
c조건	0.0625 (0.0019)	0.8757 (0.0127)
d조건	0.0673 (0.0019)	0.8690 (0.0143)



[그림 Ⅳ-5] 피험자 능력분포가 정규분포를 이룰 때
배점 조건에 따른 일반화가능도 계수의 변화

3. 피험자 분포가 부적편포인 경우 신뢰도의 변화

피험자 분포가 위와 달리 부적편포를 이룬다고 가정했을 때의 여러 배점 방식에 따라 분산성분 추정치를 분석한 결과는 <표 Ⅳ-7>, <표 Ⅳ-8>과 같다. 동등배점 조건에서 오차요인의 영향력은 피험자에 의한 분산성분 추정치가 약 81.7~81.9%로, 이는 피험자 분포가 정규분포일 때보다 약간 낮은 수치였다. 피험자와 문항의 상호작용에 의한 분산성분 추정치는 약 13.7~13.8%, 문항에 의한 분산성분 추정치는 약 4.5%로, 이는 정규분포를 가정하였을 때에 비하여 약간 높은 수치였다.

차등배점 조건에서 오차요인의 영향력은 피험자와 문항의 상호작용에 의한 분산성분 추정치가 약 71.3~78.9%로 가장 높았다. 피험자에 의한 분산성분 추정치는 약 10.1~27.3%로 다음으로 높았으며, 문항의 차이에 의한 분산성분 추정치는 약 1.3~11.1%로 가장 낮았다.

이를 바탕으로 30문항에 대하여 D연구의 오차분산과 신뢰도 계수를 산출한 결과를 제시하면 <표 Ⅳ-9>, <표 Ⅳ-10>과 같다. 피험자의 능력분포가 부적편포를 이룰 때, 동등배점을 적용했을 때의 일반화가능도 계수는 약 .85로 양호한 신뢰도로 나타났다. 차등배점을 적용했을 경우 차등배점의 간격이 커질수록(b조건 → d조건) 신뢰도가 약 .85에서 .82로 줄어들었으나 모두 .80이상의 신뢰도를 만족하여 검사의 신뢰도는 모두 양호한 수준으로 나타났다. 이를 그래프로 나타내면 [그림 Ⅳ-6]과 같다.

<표 Ⅳ-7> 100명일 때 차등배점 조건에서 G연구의 분산성분 추정치

(피험자 분포가 부적편포를 이룰 경우)

(평균, 비율)

조건	배점	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(p,i)$
a조건	동등배점	0.3464 (81.83%)	0.0188 (4.45%)	0.0581 (13.72%)
b조건	v1 (저배점)	0.1939 (26.81%)	0.0098 (1.36%)	0.5194 (71.83%)
	v2 (중배점)	0.4822 (22.86%)	0.0798 (3.78%)	1.5475 (73.36%)
	v3 (고배점)	0.3152 (10.39%)	0.3373 (11.11%)	2.3822 (78.50%)
c조건	v1 (저배점)	0.0963 (26.22%)	0.0059 (1.61%)	0.2649 (72.17%)
	v2 (중배점)	0.4231 (22.85%)	0.0688 (3.72%)	1.3592 (73.43%)
	v3 (고배점)	0.3856 (10.30%)	0.4139 (11.05%)	2.9442 (78.64%)
d조건	v1 (저배점)	0.0359 (27.26%)	0.0019 (1.44%)	0.0940 (71.31%)
	v2 (중배점)	0.3653 (22.65%)	0.0606 (3.76%)	1.1868 (73.59%)
	v3 (고배점)	0.4643 (10.25%)	0.5020 (11.08%)	3.5647 (78.67%)

<표 Ⅳ-8> 300명일 때 차등배점 조건에서 G연구의 분산성분 추정치

(피험자 분포가 부적편포를 이룰 경우)

(평균, 비율)

조건	배점	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(p,i)$
a조건	동등배점	0.3438 (81.70%)	0.0191 (4.53%)	0.0580 (13.77%)
b조건	v1 (저배점)	0.1885 (26.86%)	0.0090 (1.28%)	0.5043 (71.86%)
	v2 (중배점)	0.4803 (22.79%)	0.0796 (3.78%)	1.5474 (73.43%)
	v3 (고배점)	0.3064 (10.13%)	0.3317 (10.97%)	2.3855 (78.90%)
c조건	v1 (저배점)	0.0974 (27.24%)	0.0048 (1.34%)	0.2553 (71.42%)
	v2 (중배점)	0.4225 (22.80%)	0.0697 (3.76%)	1.3608 (73.43%)
	v3 (고배점)	0.3813 (10.22%)	0.4099 (10.99%)	2.9395 (78.79%)
d조건	v1 (저배점)	0.0351 (27.21%)	0.0017 (1.33%)	0.0921 (71.45%)
	v2 (중배점)	0.3702 (22.92%)	0.0610 (3.78%)	1.1842 (73.31%)
	v3 (고배점)	0.4598 (10.19%)	0.4946 (10.96%)	3.5564 (78.84%)

<표 Ⅳ-9> 100명일 때 차등배점 조건에서 D연구의
오차분산 및 신뢰도 계수
(피험자 분포가 부적편포를 이룰 경우)

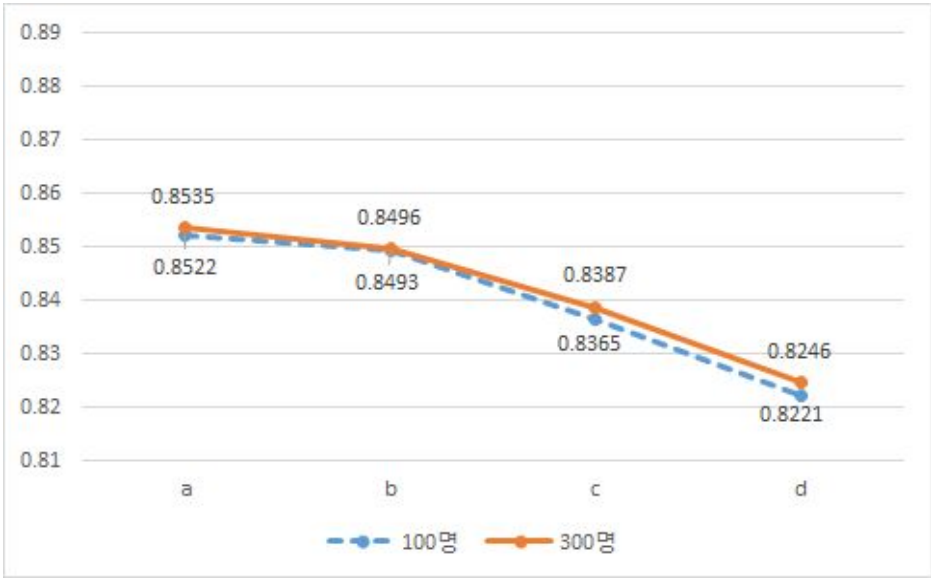
(평균, 표준편차)

배점조건	상대오차분산	일반화가능도 계수
a조건	0.0581 (0.0034)	0.8522 (0.0268)
b조건	0.0602 (0.0033)	0.8493 (0.0271)
c조건	0.0646 (0.0031)	0.8365 (0.0305)
d조건	0.0703 (0.0031)	0.8221 (0.0333)

<표 Ⅳ-10> 300명일 때 차등배점 조건에서 D연구의
오차분산 및 신뢰도 계수
(피험자 분포가 부적편포를 이룰 경우)

(평균, 표준편차)

배점조건	상대오차분산	일반화가능도 계수
a조건	0.0580 (0.0031)	0.8535 (0.0193)
b조건	0.0602 (0.0029)	0.8496 (0.0201)
c조건	0.0645 (0.0027)	0.8387 (0.0216)
d조건	0.0701 (0.0026)	0.8246 (0.0247)



[그림 Ⅳ-6] 피험자 능력분포가 정규분포를 이룰 때
배점 조건에 따른 일반화가능도 계수의 변화

4. 문항 수 조건에 따른 신뢰도의 변화

일반화가능도 분석의 D연구를 통해 문항의 수를 20문항과 25문항으로 변화시켜 일반화가능도 계수의 변화를 살펴본 결과는 <표 Ⅳ-11>에서 <표 Ⅳ-14>까지와 같다. 피험자 능력 분포가 정규분포를 따를 경우, 전반적으로 문항 수가 적고 차등 배점 간격이 커질수록 신뢰도가 낮아지는 경향은 있었으나, 가장 낮은 경우에도 .81이상의 신뢰도를 만족하였다. 그러나 피험자 능력 분포가 부적편포를 따른다고 가정했을 경우, 문항 수를 25문항으로 줄이면 배점 간격이 가장 큰 조건(d조건)에서 신뢰도가 약 .79로 나타났다. 또한, 문항 수를 20문항으로 더욱 줄일 경우, 검사의

신뢰도는 약 .75~.80의 분포를 보여 상당수가 양호한 수준의 신뢰도에 미치지 못하는 것으로 나타났다.

<표 Ⅳ-11> 정규분포, 100명일 때 문항 수 조건에 따른 신뢰도 계수

(평균, 표준편차)

문항 수	a조건 (동등배점)	b조건	c조건	d조건
30문항	0.8780 (0.0180)	0.8779 (0.0189)	0.8731 (0.0196)	0.8660 (0.0208)
25문항	0.8572 (0.0206)	0.8551 (0.0220)	0.8494 (0.0228)	0.8408 (0.0243)
20문항	0.8278 (0.0239)	0.8266 (0.0253)	0.8204 (0.0262)	0.8111 (0.0278)

<표 Ⅳ-12> 정규분포, 300명일 때 문항 수 조건에 따른 신뢰도 계수

(평균, 표준편차)

문항 수	a조건 (동등배점)	b조건	c조건	d조건
30문항	0.8802 (0.0116)	0.8803 (0.0117)	0.8757 (0.0127)	0.8690 (0.0143)
25문항	0.8596 (0.0133)	0.8578 (0.0138)	0.8523 (0.0149)	0.8443 (0.0168)
20문항	0.8305 (0.0155)	0.8298 (0.0159)	0.8238 (0.0172)	0.8150 (0.0193)

<표 Ⅳ-13> 부적편포, 100명일 때 문항 수 조건에 따른 신뢰도 계수

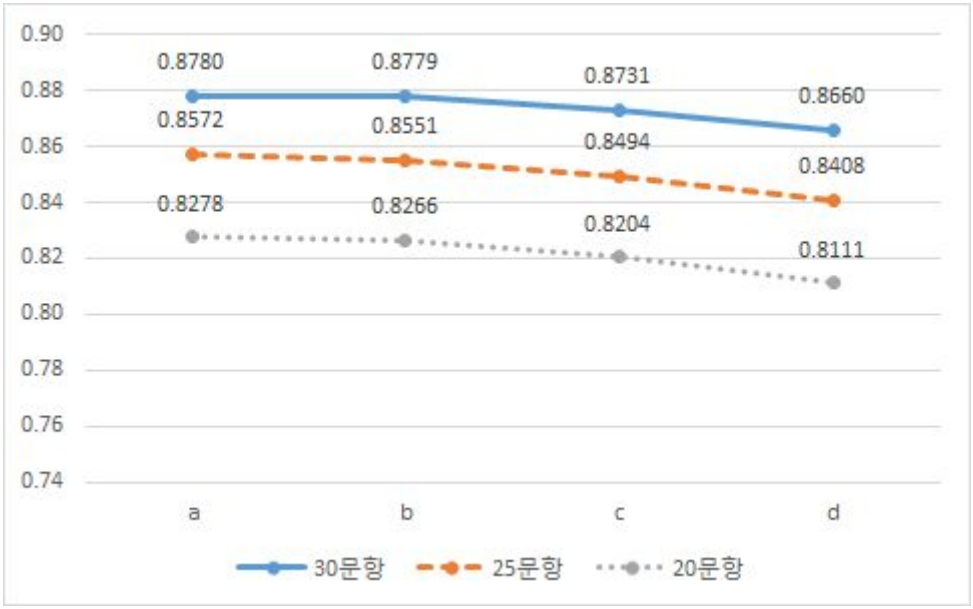
(평균, 표준편차)

문항 수	a조건 (동등배점)	b조건	c조건	d조건
30문항	0.8522 (0.0268)	0.8493 (0.0271)	0.8365 (0.0305)	0.8221 (0.0333)
25문항	0.8278 (0.0303)	0.8221 (0.0313)	0.8068 (0.0350)	0.7898 (0.0381)
20문항	0.7939 (0.0346)	0.7874 (0.0356)	0.7709 (0.0395)	0.7528 (0.0427)

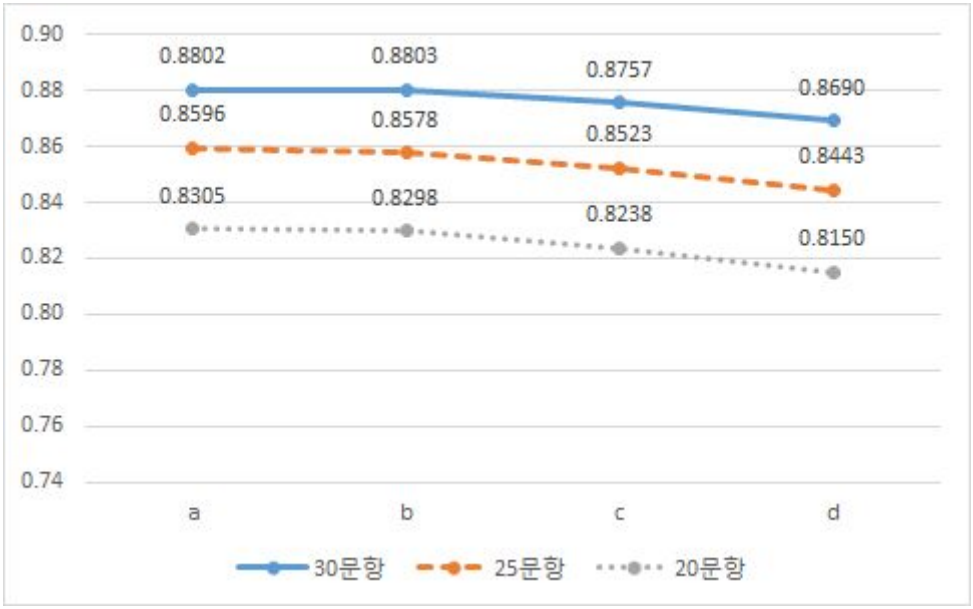
<표 Ⅳ-14> 부적편포, 300명일 때 문항 수 조건에 따른 신뢰도 계수

(평균, 표준편차)

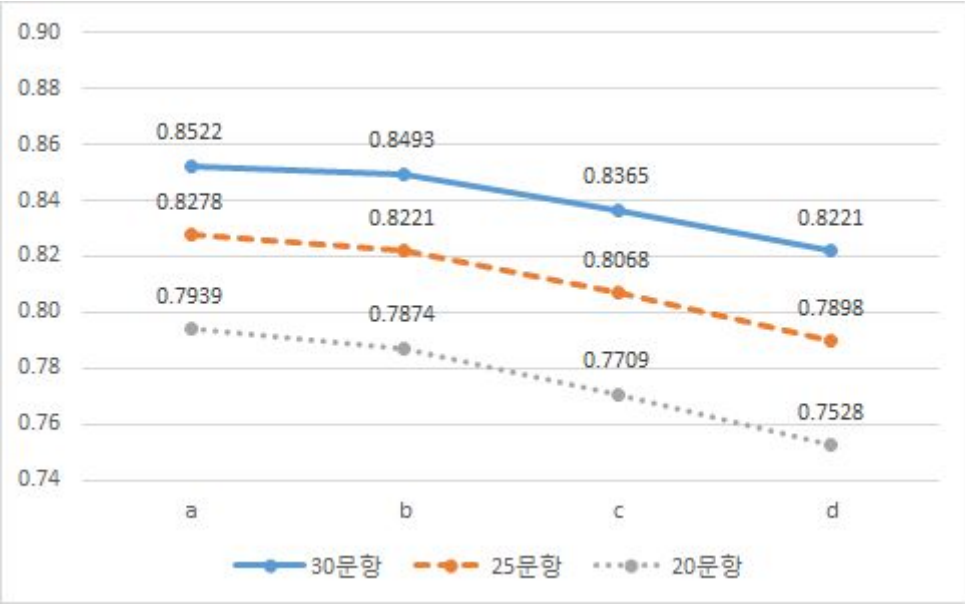
문항 수	a조건 (동등배점)	b조건	c조건	d조건
30문항	0.8535 (0.0193)	0.8496 (0.0201)	0.8387 (0.0216)	0.8246 (0.0247)
25문항	0.8293 (0.0219)	0.8224 (0.0233)	0.8093 (0.0250)	0.7924 (0.0285)
20문항	0.7955 (0.0251)	0.7876 (0.0268)	0.7735 (0.0284)	0.7556 (0.0321)



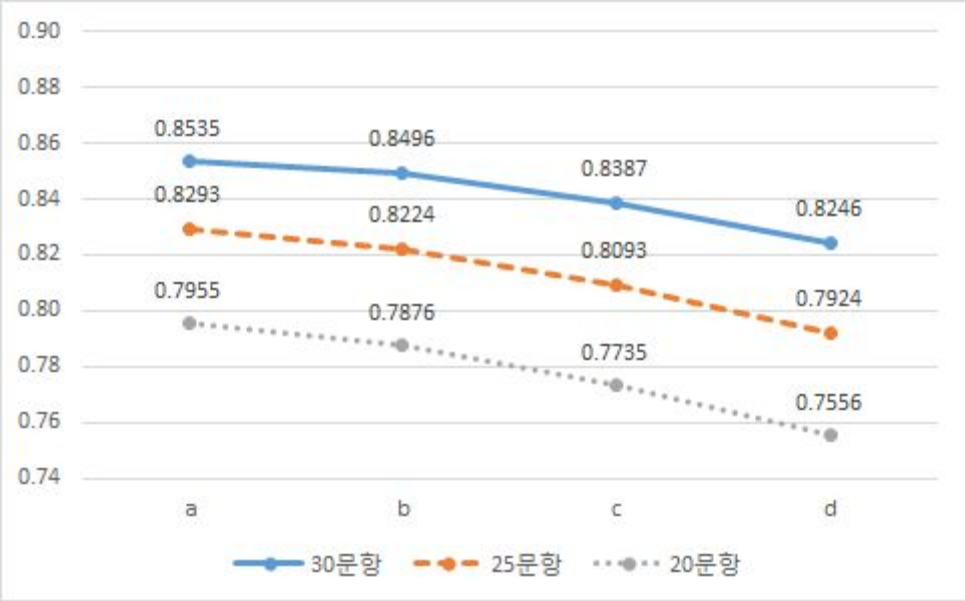
[그림 Ⅳ-7] 정규분포, 100명일 때 문항 수 조건에 따른 신뢰도 변화



[그림 Ⅳ-8] 정규분포, 300명일 때 문항 수 조건에 따른 신뢰도 변화



[그림 Ⅳ-9] 부적편포, 100명일 때 문항 수 조건에 따른 신뢰도 변화



[그림 Ⅳ-10] 부적편포, 100명일 때 문항 수 조건에 따른 신뢰도 변화

Ⅴ. 요약 및 논의

1. 요약

우리 주변에 있는 수많은 검사 중 검사를 구성하는 여러 요소에 따라 검사의 측정학적인 타당성은 크게 달라질 수 있다. 사전에 설정된 성취기준에 의해 평가가 이루어지는 준거참조평가가 시간이 지나자 본질이 왜곡된 임의적 평가라는 비판을 받을 때도 있고(김신영, 백순근, 채선희, 1998), 검사의 신뢰도와 관련하여 충분한 분석 절차 없이 정해진 가중치가 검사의 신뢰도를 보장하지 못하는 경우도 있으며(양길석, 2007; 김주훈 외, 2010; 이현숙, 2012), 검사 제작 시에 당연시했던 가정이 실제 검사 결과에서는 틀린 것으로 나타날 수도 있다(이영식, 신상근, 2004). 그렇기 때문에 평가 결과에 대한 해석이 타당한가에 대해 점검해보는 노력이 계속해서 이루어져야 하며(이영식, 신상근, 2004), 특히 민감하고 중요한 검사 상황에서는 검사점수의 타당성과 신뢰성을 확보하기 위한 더 많은 노력이 필요하다.

이 연구는 중・고등학교에서 이루어지는 검사의 점수가 교육측정학적 관점에서 타당하고 신뢰도가 높은지에 대해 의문을 제기한 것으로부터 시작하였다. 이에 따라 이 연구의 목적은 학교에서 이루어지는 혼합형 검사의 신뢰도가 검사 문항과 배점 방식에 따라 달라질 수 있음을 고려하여, 검사를 구성하는 문항의 가중치와 피험자의 분포, 문항 수가 변함에 따라 신뢰도가 어떻게 달라지는지 분석하는 것이었다. 문항 수와 검사를 구성하는 각 문항이 가중치는 전문가에 의해 내용적으로 그 타당성

을 확인하는 부분이 크지만, 측정학적 측면에서 검사의 신뢰도를 만족하지 못하면 타당도 또한 높을 수 없다는 점에서(성태제 ,김경희, 1993) 검사의 측정학적 신뢰도를 분석하고자 하였다.

이를 위해 문항의 배점 방식을 오차요인 중 하나로 설정하여 다변량 일반화가능도 모형을 설계하고 적용하였으며, 대학수학능력시험의 수리영역의 틀을 빌려 혼합형 검사에서 문항 배점 조건 및 피험자 규모에 따른 신뢰도의 변화를 살펴보고자 하였다. 모의실험을 위한 자료를 생성하였는데 그 기준은 피험자 규모(100명, 300명), 배점 조건(동등배점 1개 조건, 차등배점 3개 조건), 피험자 분포(정규분포, 부적편포), 문항 수(30문항, 25문항, 20문항)가 있었다. 피험자 규모를 바탕으로 자료를 생성한 후 피험자의 문항 정답 확률과 문항반응자료를 만들었고, 이에 일반화가능도 이론을 적용하여 배점 조건에 따라 검사를 구성하는 오차요인의 분산성분을 추정하고 적정한 신뢰도를 만족시킬 수 있는 검사 조건을 알아내고자 하였다. 동등배점을 가정한 경우 단일국면설계인 $p\times i$ 설계를, 차등배점을 가정한 경우 다변량 일반화가능도 설계인 $p^{\bullet}\times i^{\circ}$ 설계를 적용하여 분석하였다. 일반화가능도 설계의 D연구에서는 문항수를 고려하여 오차분산과 신뢰도 계수를 산출하였다.

이에 따라 분석한 연구결과를 요약하면 다음과 같다.

첫째, 분산성분 추정치를 통해 동등배점인지 차등배점 방식인지에 따라 피험자, 문항, 상호작용의 영향력이 다르게 나타났다. 동등배점인 경우 피험자의 능력 차이에 의한 분산성분이 가장 크고, 문항에 의한 분산성분이 가장 작게 나타났다. 반면 차등배점을 가정한 설계에서는 피험자와 문항의 상호작용에 의한 분산성분이 가장 크고, 문항에 의한 분산성분이 가장 작게 나타났다.

둘째, 동등배점 조건에서 검사의 신뢰도는 차등배점 조건에서 검사의

신뢰도보다 일반적으로 높게 나타났다. 피험자 수, 피험자 능력 분포와 상관없이 동등배점 조건에서의 일반화가능도 계수는 모두 .85 이상으로 양호한 수준의 신뢰도를 만족하였다.

셋째, 차등배점 조건을 적용했을 때 검사의 신뢰도는 배점 간격이 커질수록 감소하였고, 피험자 능력 분포가 부적편포를 이룰 때 그 감소 정도는 더 크게 나타났다. 피험자 능력 분포가 정규분포를 이루는 경우 차등배점 조건에 따른 신뢰도의 변화는 약 .88에서 .87로 감소하였고, 피험자 능력 분포가 부적편포를 이루는 경우 차등배점 조건에 따라 신뢰도의 변화는 약 .85에서 .82로 감소하였다.

넷째, 문항의 수가 줄어들수록 신뢰도는 비교적 크게 감소하였다. 피험자 능력 분포를 정규분포로 가정한 경우, 검사 문항을 30문항에서 25문항으로 줄이면 신뢰도는 약 .87～.88 수준에서 약 .84～.86 수준으로 감소하였다. 검사 문항이 20문항인 경우 신뢰도는 약 .81～.83 정도인 것으로 나타났다. 피험자 능력 분포가 부적편포임을 가정한 경우, 30문항일 때의 신뢰도는 약 .82～.85였으나, 검사 문항이 25문항일 때 신뢰도는 약 .79～.83 수준이었다. 검사 문항이 20문항인 경우 신뢰도는 약 .75～.80 정도를 나타내어 양호한 수준의 신뢰도 .80을 넘지 못하는 경우도 있었다.

2. 논의

연구결과를 바탕으로 논의할 점은 다음과 같다.

첫째, 동등배점을 적용한 검사의 신뢰도가 차등배점을 적용한 검사의 신뢰도보다 일반적으로 비교적 높게 나타났다. 이는 차등배점을 적용할 특별한 이유와 근거가 없는 한 동등배점을 사용하는 것이 타당하다는 선행연구(김신영, 노국향, 1999; Russell 외, 2006; 양길석, 2007; 김주훈 외, 2010)의 결과를 뒷받침한다. 차등배점의 여러 조건에 따라서는 배점 간 간격이 클수록 신뢰도가 감소하는 것으로 나타났다. 그러나 문항 수가 30문항일 경우, 피험자 수와 피험자 능력 분포, 배점 조건에 상관없이 모든 경우 .80 이상의 신뢰도를 만족하였다. 이는 일반화가능도 이론을 적용하는 많은 연구에서 적정수준의 일반화가능도 계수를 .80으로 설정하였음(김성숙, 1993; 김성숙, 1995; Nocera, Ferlazzo, & Borghi, 2001; 이선영 외, 2015)을 고려했을 때, 문항 수가 확보된다면 배점 조건은 검사의 신뢰도에 크게 영향을 주지 않는다고 볼 수 있다.

둘째, 차등배점을 적용할 경우 문항 간 배점 차이가 신뢰도의 감소에 미치는 영향은 비교적 선형적이었다. 이는 [그림 Ⅳ-5], [그림 Ⅳ-6]에서 시각적으로 확인할 수 있는데, 차등배점의 간격이 커질수록 신뢰도가 꾸준히 감소하였다. 이는 이 연구에서 차등배점 조건을 고려할 때 대학수학능력시험 수리영역의 문항 구성 틀을 바탕으로 차등배점의 가지 수나 고배점과 중배점, 중배점과 저배점 사이의 점수 차이를 일정하게 고정하였기 때문인 것으로 보인다. 따라서 문항의 수를 조절하면서 배점 간 간격과 배점의 가지 수를 조금 더 다양하고 뚜렷하게 설정할 경우 그 방법에 따라 차등배점이 신뢰도에 미치는 영향은 다르게 나타날 수 있다.

셋째, 문항 수에 따라 검사의 신뢰도는 비교적 크게 변화하였는데 검

사의 신뢰도가 양호한 수준인 .80 이상을 만족하기 위해서는 피험자 능력 분포가 정규분포를 이루거나, 부적편포가 있는 경우 문항 수가 적어도 25문항 이상이 되어야 함을 알 수 있었다. 피험자 능력 분포가 정규분포를 이루고 있다고 하더라도 검사 문항이 20문항인 경우, 차등 배점 간격이 커지면(d조건) 피험자 수가 100명인 경우에 신뢰도 평균이 약 .81, 표준편차가 약 .03 정도로 감소하게 된다. 이러한 점을 고려했을 때, 차등 배점을 적용할 경우 가장 높은 배점과 가장 낮은 배점 간의 차이가 신뢰도에 영향을 줄 만큼 커지지 않도록 설정하는 것이 필요하며, 이는 가장 높은 배점과 가장 낮은 배점 간 차이가 3배 미만이 되어야 이해 당사자들 간의 논란을 최소화 할 수 있다는 김주훈 외(2010)의 연구를 뒷받침한다고 볼 수 있다.

이 연구에서는 차등배점을 고려한 학력검사에서 차등배점과 피험자 분포, 문항 수의 조건에 따라 검사의 신뢰도를 일반화가능도 이론을 적용하여 분석하였다. 이 연구는 혼합형 검사를 구성하는 문항의 가중치를 일반화가능도 이론의 한 국면으로 설정하고 이 조건에 따라 변화하는 신뢰도를 살펴보았다는 점에서 의의가 있으며, 전술한 연구 결과는 적정한 신뢰도를 유지할 수 있는 과목별 배점 방식과 문항 수를 예시하거나 혹은 어떤 평가를 구성하는 세부 검사들의 가중치를 고려하여 평가의 신뢰도를 향상시키는 데에도 활용할 수 있을 것으로 기대된다.

이 연구의 제한점을 살펴보면 다음과 같다.

첫째, 실제 자료가 아닌 모의실험 설계로 생성된 자료를 사용하였다. 이는 개별 학교 현장에서 쓰이는 검사 결과를 활용하기 어렵고 차등배점을 고려한 선행연구가 부족했던 것을 고려해 모의실험을 설계하였다 하더라도 실제 검사에서 존재했을 오차요인이나 자료의 분포를 고려하지 못했다는 점에서, 또 대학수학능력시험을 바탕으로 한 시험의 문항 구성

이 개별 학교들의 시험 구성과 차이가 있을 수 있다는 점에서 한계가 있다.

둘째, 다양한 차등배점 요인을 고려하지 못하였다. 이 연구에서는 특정한 시험의 구조를 변형하여 차등배점 요인을 3가지 배점 방식과 선형적인 차이로만 고정하였다. 그러나 학교 현장에서는 검사 문항의 가중치가 비선형적으로 구성되거나 3가지 보다 많은 종류의 배점방식을 갖고 있을 수 있다는 점에서 분석결과를 일반화시킬 수 없다는 한계가 있다.

후속연구에 대한 제언으로 앞서 서술한 제한점들을 고려하여 실제 자료를 바탕으로 배점 조건을 포함한 일반화가능도 분석을 실시할 것을 생각해 볼 수 있을 것이다. 또한 차등배점 방식의 다양화에 따른 신뢰도의 변화에 대해서도 분석한다면 배점 조건에 따른 학력평가의 신뢰도를 보다 정확히 분석하는데 도움이 될 것이라 생각한다.

참 고 문 헌

김경선, 이규민, 강승혜(2010). 일반화가능도 이론을 적용한 한국어 말하기 성취도 평가의 신뢰도와 오차요인 분석. **한국어 교육, 21**(4), 51-75.

김도연, 허종관(2002). 일반화가능도이론을 적용한 주관적 배구기능검사의 신뢰도 추정. **한국체육측정평가학회지, 4**(2), 15-28.

김명화(2005). 채점 자동화 시스템 구축을 위한 수학 구성형 문항 채점의 일반화 가능도 연구. **교육문제연구, 22**, 205-222.

김성숙(1992). 관찰체계에 있어 측정의 변동요인 분석-관찰자 합치도, 안정도, 일반화가능도 비교. **교육평가연구, 5**(1), 37-56.

김성숙(1993). 관찰을 통한 교수 평가 체계에 대한 측정의 일반화 가능도 연구. **교육학연구, 31**(1), 23-40.

김성숙(1995). 논술 문항 채점의 변동 요인 분석과 일반화가능도계수의 최적화 조건. **교육평가연구, 8**(1), 35-57.

김성연, 한기순(2013). 관찰· 추천제에 의한 수학영재 선발 시 사용되는 교사추천서와자기소개서 평가에 대한 다변량 일반화가능도 이론의 활용. **영재교육연구, 23**(5).

김신영, 노국향(1999). 선택형 검사문항에 대한 차등배점의 타당성에 관한 연구. **교육평가연구, 12**(2), 235-250.

김신영, 백순근, 채선희(1998). 국가 수준의 '성취기준 및 평가기준'개발에 대한 고찰. **교육평가연구, 11**(1), 47-73.

김정환, 이용환(1999). 2-단계 맞춤검사에서 일반화가능도 계수의 최적화 조건. **교육평가연구, 12**(1), 65-82.

김주훈, 동효관, 송미영, 남민우, 김미영, 최원호, 이재봉, 주지은, 이은경, 김지윤(2010). 문항의 배점 결정 요인 및 타당성 분석 - 2009학년도 의학교육입문검사를 중심으로. **교육과정평가연구, 13**(2), 197-218.

노국향(1995). 행렬표집(matrix sampling) 설계를 이용한 학교 작문 검사의 일반화가능도에 관한 연구. **교육평가연구, 8**(1), 105-122.

노국향, 박정(2001). 문항의 형태와 배점에 따른 검사 정보의 비교. **교육평가연구, 14**(2), 173-191.

박찬호, 강태훈(2011). 전문가 판정에 의한 차등 배점을 활용한 제한적 일반화부분점수 모형의 적용. **교육평가연구, 24**(3), 781-797.

서울시교육청(2015a). **중학교 학업성적관리 시행지침**. 서울특별시 교육청.

서울시교육청(2015b). **고등학교 학업성적관리 시행지침**. 서울특별시 교육청.

성태제, 김경희(1993). 문항수, 문항난이도, 문항변별도 변화에 따른 신뢰도 계수와 검사정보함수의 변화. **교육평가연구, 6**(2), 123-154.

신동일(2001). 일반화가능도 이론 적용을 중심으로 한 말하기 평가도구 타당도 검증 연구. **응용언어학, 17**(1), 199-221.

신택수(2016). 잠재혼합효과 분석방법의 수행력 비교연구. **교육평가연구, 29**(1), 51-77.

양길석(2007). 전문가 판단방법에 따른 문항 간 차등배점의 타당성 연구. **한국교육학연구, 13**(1), 197-219.

양지승, 이규민(2007). 일반화가능도 이론을 적용한단위검사 구성 검사점수의 신뢰도 추정. **교육평가연구, 20**(1), 9-139.

이기영, 안희수(2005). 중등학교 과학 수행평가의 평가 유형과 채점 방식

및 신뢰도 분석. **한국과학교육학회지, 25**(2), 173-183.

이문수, 이규민, 강상진(2009). 모의 실험을 통한 혼합형 문항 검사의 문항반응이론 척도변환과 진점수동등화 조건 탐색. **교육평가연구, 22**(3), 805-826.

이선영, 김성연, 김정하, 백근찬, 이병윤(2015). 다변량 일반화가능도 이론을 활용한 창의성 예비검사의 신뢰도 분석. **창의력교육연구, 15**(3), 83-107.

이영식, 신상근(2004). 다변량 일반화가능도 이론에 의한 말하기 시험의 타당도와 신뢰도에 관한 연구. **외국어교육, 11**(2), 249-265.

이은하(2015). 한국어 쓰기 수행 평가 신뢰도 추정 연구-일반화가능도 이론, 알파 계수 및 채점자 사후 설문조사를 사용하여. **국어교육, 149**(단일호), 241-277.

이종성(1997). 일반화가능도 이론의 연구과제. **연세교육과학, 45**, 1-15.

이현숙(2012). 혼합형 검사의 문항 유형별 가중치에 따른 신뢰도 및 다변량 일반화가능도 분석. **교육평가연구, 25**(1), 95-116.

임인재, 김신영, 박현정(2011). **심리측정의 원리**. 서울: 화연사.

한국교육과정평가원(2016. 3. 29.). **2017학년도 대학수학능력시험 시행 기본계획**. 한국교육과정평가원.

Bayuk, R. J. (1973). The effects of choice weights and item weights on the reliability and predictive validity of reading tests. ***In annual meeting of the American Educational Research Association****, Chicago, Illinois.*

Brennan, R. L. (1992). Generalizability theory. ***Educational Measurement: Issues and Practice, 11***(4), 27-34.

Brennan, R. L. (2001). ***Generalizability theory: Statistics for social***

science and public policy. New York: Springer-Verlag. Retrieved March, 30, 2013.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44*(2), 131-155.

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

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological methods, 1*(1), 16.

Di Nocera, F., Ferlazzo, F., & Borghi, V. (2001). G theory and the reliability of psychophysiological measures: A tutorial. *Psychophysiology, 38*(05), 796-806.

Eggen, T. J., & Veldkamp, B. P. (2012). *Psychometrics in Practice at RCEC*. Ipskamp Drukkers.

Kim, S., & Lee, W. C. (2004). IRT Scale Linking Methods for Mixed-Format Tests. *ACT Research Report Series 2004-5*. ACT Inc.

Koizumi, R. (2015). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, 0265532215587390.

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating

schemes. *ETS Research Report Series, 2005*(1), i-76.

Powers, S., & Brennan, R. L. (2009). Multivariate generalizability analyses of mixedformat exams. *In Annual Meeting of the National Council on Measurement in Education*, San Diego, CA.

Russell, L. B., Hubley, A. M., Palepu, A., & Zumbo, B. D. (2006). Does weighting capture what's important Revisiting subjective importance weighting with a quality of life measure. *Social Indicators Research, 75*(1), 141-167.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series, 2006*(1), i-71.

Abstract

Generalizability Analysis of Student Achievement Tests with Various Item Weights

Shinhye Lee
Department of Education
The Graduate School
Seoul National University

Mixed format tests with various item weights are mainly used at middle and high schools in Korea to improve discrimination of student's ability. This is to keep students from having equal scores when grading is necessary, and also to increase validity by

considering item difficulty and given time to solve problems. In terms of educational measurement, however, tests with various item weights can intimidate test's reliability, thereby threatening test validity, especially because test reliability depends on the test components.

Research about item weights related to test reliability and validity has been conducted continuously. Item weights used in school fields are usually decided by the subject expert according to item difficulty and importance. Expert-generated item weights can be useful if experts judge the characteristics of examinee and items correctly, because they help discriminate the examinee and provide proper test results. However, when the criterion of judgement is not clear, the test can be criticized for subjectiveness and randomness. For these reasons, the appropriateness of item weights are controversial in many studies.

This study analyzed and compared test reliability according to several conditions including various item weights, examinee distribution, and the number of test items. For this study, simulation data is generated and analyzed using generalizability theory. Simulation data follows the form of College Scholastic Ability Test(CSAT) in Korea and has three generating conditions such as item weights, size of examinee, and distribution of examinee. After 500 times iteration, the average reliability could be calculated, and in the process, the reliability depending on the number of items could be also calculated.

The result of this study is as follows. First, test reliability using

differential item weights is generally lower than the reliability of tests with equally weighted items. Similar to preceding research, this shows that differential item weights are no better than equal item weights unless there is definite reason to use various item weights. Second, the test reliability decreased as the range of item weights increased. This shows that it is important to use proper range of item weights for better reliability. Third, the test reliability decreased relatively as the number of items reduced. Especially, when the distribution of examinee was negatively skewed and the number of items was 20, the reliability was below than .80, which shows that when the examinee are distributed with skewness, the test items should consist of more than 25 items to keep appropriate reliability.

This study analyzed test reliability with various item weights, two examinee distributions, and different number of items using generalizability theory. In this regard, this study illustrates efficient item weights and the number of items for stable reliability. In addition, this study can be applied to analyze the reliability of assessment consisting of various sub-tests.

The limitation of this study is that because it analyzed simulation data, error factors such as different examinee distribution were beyond consideration. In this study, also, differential item weights are adjusted using only linear variation although the weight variation is not linear in school fields. For a follow-up study, therefore, it is suggested to analyze test reliability with differential item weights using real data as well as with consideration of additional error

factors and variation of item weights.

Keywords : reliability, item weights, generalizability theory, simulation study
Student Number : 2014-20841