韓國數學敎育學會誌 시리즈 A <數學敎育>
https://doi.org/10.7468/mathedu.2018.57.1.55
2018. 02. 제 57권, 제 1호, 55-74.

J. Korean Soc. Math. Ed. Ser. A:
The Mathematical Education
February. 2018, Vol. 57, No. 1, 55-74.

예비수학교사의 교직 적성·인성 검사에서 분할점수 변화에 따른 다양한 신뢰도 탐색

김 성 연 (인천대학교)

Ⅰ. 서론

우리나라는 2013년부터 학교폭력 등 다양한 교실 상황에 적절히 대응할 수 있는 교직적성과 인성을 갖춘 교사를 선발하기 위해 모든 교원양성기관에서 2회 이상 교직 적성·인성 검사 실시를 의무화하고 있다(교육과학기술부, 2012). 이에 교육부에서는 교원양성기관에서 자율적으로 활용할 수 있도록 교직 적성·인성 검사 표준안(김정환 외, 2012)을 보급하였다. 교육대학교, 사범대학교, 교육대학원에 재학 중이거나 교직과를 이수하는 예비교사들은 재학 중인 교원양성기관의 장이 실시하는 교직 적성·인성 검사에서 합격 기준 이상의 적격 판정을 받아야만 '교원자격검정령 제19조 제3항'에 의해 교원 자격증을 취득할 수 있다(교육부, 2013). 이와 같은 자격 검사의 목적은 무능한 교사에게서 공공을 보호하기 위한 것으로(Phillips & Camara, 2006), 검사의 형태는 능력 차이를 구분해서 가장 우수한 사람을 선발하기 위해 만들어진 규준참조검사(norm-referenced test)와 달리 모든 자격이 최소한의 지식과 기술을 확신할 수 있도록 만들어진 준거참조검사(criterion-referenced test)에 해당한다. 따라서 검사 결과는 적격 및 부적격의 두 가지 형태로 보고되므로, 예비교사들의 교직 적성 및 인성 수준을 구분하기 위한 한 개의 분할점수(cut scores)를 필요로 한다.

분할점수는 척도 점수 위에 설정되는 특정 점수를, 수준(standard)은 피험자가 알고 있거나 할 수 있는 역량 정도를, 그리고 수준설정(standard setting) 또는 준거설정(criterion setting)은 적격 수준에 도달하기 위해 필요한 최소한의 기준점수를 정하는 과정을 말한다(Kane, 2001). 교직 적성·인성 검사 표준안(김정환 외, 2012)에서는 준거설정 절차로 변형된 Angoff 방법을 적용하였으며, 전문가, 경력 교사, 그리고 연구자들에게 문항이 아닌 하위영역에 대하여 교직 적성 및 인성의 자질을 갖춘 학생들이 취득할 수 있는 최저점수를 판정하게 하였다. 구체적으로 840점 만점에 6명으로 구성된 전문가 집단에서 설정한 분할점수의 최솟값은 545점이며, 최댓값은 609점으로 평균 572.67점이며, 비슷하게 65명으로 구성된 경력 교사 집단에서 설정한 분할점수의 평균은 570.72점으로 제시하고 있다. 반면에 검사 개발자 4명과 연구생 8명이 포함된 12명으로 구성된 연구자 집단에서 설정한 분할점수는 598점으로 다른 집단보다 다소 높게 설정되었는데, 이는 유능한 교사로서 자질을 갖춘 신임 교사가 갖추어야 할 교직 적성 및 인성 특성 때문이라고 밝히고 있다. 또한 교직 적성·인성 검사의 분할점수는 사용하는 기관의 특성에 따라 해당 기관의 검사 도구 개발 위원회에서 충분히 논의하고 합의하여 결정할 수 있어야 하지만, 국가수준의 분할점수는 연구자 집단에서 설정한 점수가 합리적이라고 제시하였다.

실제 교원양성기관 현장에서는 교육부에서 보급한 동일한 검사 표준안을 사용하는 경우에도 총점을 기준으로, 하위영역을 기준으로, 또는 총점의 기준과 동시에 몇 개의 하위영역에 대해 다양한 분할점수를 사용하고 있다. 검사점수를 산출하는 경우에도 검사 표준안에 제시되어 있는 답변일관성 문항들의 응답을 고려하지 않거나, 1점부터 4점까지의 응답을 0점부터 4점까지 재코딩을 하거나, 또는 이 문항들의 합만을 따로 산출하고 있다. 일례로 인천대학교와 배화여자대학교는 504점 이

* 접수일(2018년 01월 15일), 수정일(2018년 02월 22일), 게재확정일(2018년 02월 22일)
* ZDM분류 : C85
* MSC2000분류 : 97C40
* 주제어 : 교직 적성·인성 검사, 의존도 계수, 일반화가능도 이론, 준거참조검사

상, 강남대학교는 518점 이상, 조선대학교는 546점 이상, 경상대학교는 1차에서 총점 571점 이상이며, 하위 영역인 심리적 안정성과 소명감·교직관에서 30점 이상, 창원대학교는 572점 이상, 부산외국어대학교는 총점은 580점 이상이며 14개 영역 중 9개 영역 이상에서 각각 42점 이상인 경우를 적격 수준으로 판정하는 분할점수로 활용하고 있다. 그러나 이러한 분할점수 설정은 검사의 측정학적 측면에 대한 경험적인 검증 없이 다소 임의적이거나 또는 전문가들의 선험적인 판단에 의해 100점 만점의 환산점수에서 60점이나 70점 이상 등 절대점수를 활용하고 있다.

그러나 이처럼 교원양성기관의 특성에 따라 다양한 분할점수를 설정하고 있음에도 불구하고, 김성호(2017)에 따르면 2017년 현재 187개 대학 3만8천204명 중 부적격으로 판정된 예비교사는 약 0.6%인 259명이라고 밝히고 있다. 구체적으로 2014년에는 5만 124명 중 1.76%인 885명이, 2015년에는 6만3천97명 중 0.88%인 559명이, 2016년에는 5만9천771명 중 0.7%인 435명이 부적격으로 판정됨으로써 2014년 이후부터 부적격 비율은 매년 낮아지고 있으며, 2017년 기준으로 부적격으로 판정된 예비교사가 한명도 없는 기관도 87.2%인 163곳이라고 밝혔다. 또한 한국교육과정평가원(2016)에 따르면 교직에 적합하지 않은 예비교사를 걸러내는 장치인 교직 적성·인성 검사의 타당도와 신뢰도에 대한 제고가 필요하다고 밝히고 있다.

비록 전문가 집단의 판단을 요구하는 검사중심의 준거설정 방법을 통해 분할점수를 설정하여 적격과 부적격을 판정하는 것은 경제적 비용이나 시간적인 제약으로 교육현장에 직접 적용하는 것이 어려울 수 있다(민경석 외, 2014). 그러나 분할점수 설정에 따라 교직에 적합하지 않은 예비교사에게 교직에 종사할 자격을 줄 수도 있으며, 교직에 적합한 예비교사에게 교직에 종사할 자격을 주지 않을 수 있음을 고려할 때, 분할점수에 대한 객관성과 결과의 일반화가능성에 대한 연구가 필요하다. 즉, 적격과 부적격을 판정하는 분할점수 결정에 어떤 요인들이 영향을 미치고 있는지에 대한 분석이 요구된다.

현재까지 일반화가능도 이론을 활용하여 복잡한 측정상황과 관련된 오차요인을 분석하고, 일반화가능한 결과를 제시하고자 규준참조검사에서 일반화가능도 계수가 적정 수준의 신뢰도에 도달하는 효율적인 측정조건을 탐색하는 연구는 다각도에서 진행되어 왔다(김경선 외, 2010; 김성숙, 1993; 김성찬 외, 2012; 류춘렬, 이용근, 2010; 이규민, 황경현, 2007; 이진아, 한기순, 2016; 이태구, 양희원, 2016; 이한준, 강민수, 2017; 주지은 외, 2007; Lee, 2002; Tindal et al., 2010). 반면에 일반화가능도 이론을 활용하여 준거참조검사에서의 신뢰도에 해당하는 의존도 계수가 적정 수준에 도달하는 효율적인 측정조건을 탐색하는 연구는 상대적으로 소수에 불과하며(김성숙, 1998; Arce & Wang, 2012; Lin & Zhang, 2014; Taylor & Pastor, 2013), 의존도 계수를 보고한 대부분의 연구들은 규준참조검사를 절대평가로 사용할 때 참고할 수 있는 신뢰도 계수로 일반화가능도 계수와 함께 제시하고 있다(강애남, 이규민, 2006; 김보라, 이규민, 2012; 김성희, 김성연, 2017; Arterberry et al., 2012; Gugiu et al., 2012; Marzano, 2002; Newton, 2010; Schoonheim-Klein et al., 2008; Solano-Flores et al., 2006; Sung et al., 2010; Volpe et al., 2009; Volpe & Briesch, 2012; Webb et al., 2000; Yang et al., 2015).

따라서 본 연구는 준거참조검사인 교직 적성·인성 검사에 일반화가능도 이론을 적용함으로써 오차요인을 분석하고, 선행 연구들과 달리 분할점수 설정에 따라 의존도 계수 외에도 다양한 신뢰도 계수들이 동시에 적정 수준에 도달할 수 있는 효율적인 측정 조건을 탐색하는 방법을 제시하고, 이 방법의 교육현장 적용 가능성을 모색하고자 한다. 특히 기존에 교직 적성·인성 검사를 개발하고 타당화하는 연구에서는 대부분 예비교사를 교육대학교, 사범대학교, 또는 교직과를 이수하는 대학생을 대상으로 주로 이루어졌다(김경령, 서은희, 2014; 김민웅, 김태훈, 2017; 김정환 외, 2012; 서경혜 외, 2013; 조운주, 2014; 조주연 외, 2004; 조주연 외, 2007; 홍기칠, 2006). 그러나 전공 및 학교급별에 따라 교직 적성 및 인성 수준에서는 차이가 나타날 수 있으며(서경혜 외, 2013; 조운주, 2014), 외국에서 교직 적정·인성 수준을 평가한 연구에서도 교사의 개인배경에 따라 수준에는 차이가 나타났다(Englehart et al., 2012; Haq et al., 2012; Kant, 2011; Tasleema & Hamid, 2012). 특히 중등교육에서 교사는 학생들에게 가장 직접적이고 많은

영향을 미치는 존재이며, 수학교사의 신념, 교육관, 교수·학습 방법 등을 통해 학생들의 수학 성취도를 높일 수 있을 뿐만 아니라 수학에 대한 관심과 흥미, 자신감 및 가치 인식과 같은 긍정적인 태도를 갖게 할 수 있다(김정환, 2004; 김현진. 2013; 박정, 2007; 이현숙, 송미영, 2015; 한혜숙, 최계현, 2011; 황혜정, 2011; Houchard, 2005; Klieme et al., 2009; Kunter et al., 2008; Thompson, 1992). 또한 수학교육에서도 인지적 측면에서의 교육과 더불어 품성과 자질, 도덕성과 같은 인성이 강조되고 있다(신준국 외, 2017; Narvaez & Nucci, 2008). 따라서 중등교육에서 수학을 담당할 예비교사가 교직 적성 및 인성을 갖추고 있는지를 평가함으로써 그 수준을 파악하고, 부족한 부분에 대해 수준을 향상시키는 것이 필요하다. 이를 위해 측정학적 이론을 바탕으로 예비수학교사의 교직 적성·인성 검사점수에 영향을 미치는 요인들의 상대적인 영향력과 다양한 신뢰도 계수를 바탕으로 각 교원양성기관에서 자율적으로 시행하고 있는 분할점수를 바탕으로 효율적인 측정 조건을 탐색하고자 한다. 본 연구에서는 상대적으로 연구가 수행되지 않았던 교육대학원의 수학교육 전공 원생들을 대상으로 분석을 수행하였다. 구체적인 연구문제는 다음과 같다.

첫째, 교직 적성·인성 검사에서 예비수학교사인 원생의 점수에 영향을 미치는 요인들의 상대적인 영향력은 어느 정도인가?

둘째, 교직 적성·인성 검사의 다양한 신뢰도 계수가 분할점수 변화에 따라 동시에 적정 수준에 도달하기 위한 효율적인 측정 조건은 무엇인가?

Ⅱ. 이론적 배경

1. 교직 적성·인성 검사의 신뢰도

일반적으로 교직 적성·인성 검사에서는 고전검사 이론을 바탕으로 한 Cronbach α 계수를 활용하여 신뢰도를 보고하고 있으며, 연구자 및 전문가들에 의한 내용타당도, 상관분석을 통한 수렴 및 판별 타당도, 탐색적 및 확인적 요인분석을 활용하여 구인타당도를 확보하고 있다. 또한 공인타당도 분석을 위해서는 고등학교 내신 성적, 수능점수, 면접 및 논술고사 점수, 대학 성적, 창의성 검사, 인성 검사들을 활용하여 상관분석을 수행하고 있다. 이 중 본 연구와 관련이 있는 국내에서 개발된 교직 적성·인성 검사에서의 신뢰도를 살펴보면 다음과 같다.

예비유아교사를 대상으로 조운주(2014)는 김정환 외(2012)가 개발한 교직 적성·인성 검사가 이들에게 타당한지를 검증한 결과, 기존 14요인 210문항에서 9요인 140문항으로 수정할 것을 제안하였다. 이처럼 수정된 예비유사교사의 교직 적성·인성 검사의 신뢰도 분석 결과 Cronbach α 계수는 기존 김정환 외(2012)에서 보고한 신뢰도인 .69-.86보다 높은 .73-.93인 것으로 높게 나타났다. 구체적으로 전체 신뢰도 계수는 .96이며, 1요인인 계획성과 성실성·책임감에서는 .91, 2요인인 소명감·교직관에서는 .92, 3요인인 언어·의사소통능력, 지도성, 문제해결력·탐구력에서는 .93, 4요인인 심리적 안정성에서는 .84, 5요인인 봉사·희생·협동성에서는 .67, 6요인인 지식·정보능력에서는 .82, 7요인인 독립성·자주성에서는 .78, 8요인인 사려성·타인존중에서는 .78, 그리고 9요인인 창의·응용력에서는 .73으로 나타났다.

예비초등교사를 대상으로 조주연 외(2004)가 2003년에 개발한 교직 적성·인성 검사(Test for Aptitude of Primary School Teacher, TAPST)는 적성·인성 관련 6개 차원과 각 차원별로 3개의 하위요인을 갖는 총 18개의 하위요인으로 구성되어 있다. 각 하위요인에는 10개의 문항이 포함되어 있으며, 이 외에 타당도 척도에 해당하는 45문항이 포함되어 총 225문항으로 구성되어 있다. 교육대학생들과 교대에 입학하려는 지원자들에게 TAPST를 실시하고 신뢰도를 분석한 결과, Cronbach α 계수는 .63-.87로 비교적 양호한 것으로 나타났다. 구체적으로 2차 연구결과, 교수능력 차원에서는 .78-.82, 생활지도 차원에서는 .74-.87, 연구능력 차원에서는 .82-.87, 창의성 차원에서는 .63-.75, 소명감 차원에서는 .73-.83, 도덕성 차원에서는 .71-.77, 그리고 타당도 차원에서는 .63-.88로 나타났다. 1차 연구 결과에서 Cronbach α 계수가 .70미만이었던 하위요인 중에서 융통성만 .63으로 .70미만이며, 공정성과 도덕성은 각각 .71과 .74로 개선된 것으로 나타났다. 한편 홍기칠(2006)은 TAPST(조주연 외, 2004)를 교육대학의 입학전형 도구로 적용할 수 있을지를 탐색하기 위해 TAPST의 타

당도 척도만 20문항으로 개발하고, 18개 하위요인에 각 10문항씩 문항진술을 동형으로 수정하여 총 200문항의 동형검사를 개발하였다. 동형검사를 신입학지원자를 대상으로 실시하고 신뢰도를 분석한 결과, Cronbach α 계수는 .63-.88로 비교적 양호한 것으로 나타났으며, 조주연 외(2007)의 연구결과와 비슷한 경향을 나타내었다. 반면에 교육대학교 4학년 학생들을 대상으로 동형검사를 실시하고 신뢰도를 분석한 결과, Cronbach α 계수는 .61-.90으로 나타났지만, 하위요인의 상황판단력, 포용심, 융통성, 인간존중, 공정성은 .70 미만인 것으로 나타났다.

예비교사인 교육대학, 사범대학 및 교직과를 이수중인 대학생을 대상으로 김정환 외(2012)는 교육부에서 개발 보급하고 있는 교직 적성·인성 검사 표준안인 14개 하위영역과 하위영역별 15문항으로 총 210문항을 개발하였다. 종합보고서에서 제시하고 있는 예비 교직 적성·인성 검사를 중심으로 살펴보면, 14개의 하위영역과 하위영역별 25문항씩 총 350문항으로 구성되어 있다. 예비 교직 적성·인성 검사의 신뢰도 분석 결과 Cronbach α 계수는 .69-.86으로 비교적 높게 나타났다. 구체적으로 하위영역인 문제해결력·탐구력에서는 .71, 판단력에서는 .77, 독립성·자주성에서는 .74, 창의·응용력에서는 .84, 심리적 안정성에서는 .82, 언어·의사소통능력에서는 .69, 지도성에서는 .83, 공감·포용력에서는 .75, 지식·정보능력에서는 .84, 봉사·희생·협동성에서는 .84, 계획성에서는 .85, 성실성·책임감에서는 .84, 소명감·교직관에서는 .86, 그리고 열정에서는 .81로 나타났다.

한편 교직 인성에만 초점을 맞추어 김경령과 서은희(2014)는 예비교사들이 교직이라는 직업을 수행하는데 필요한 인성을 갖추었는지를 스스로 확인할 수 있는 교직 인성 자기점검도구를 개발하였다. 이 검사는 자기조절, 사회성, 도덕성의 하위요인을 갖는 보편적 인성 영역과 소명의식, 학생에 대한 열정, 교육적 신념의 하위요인을 갖는 교직 인성 영역의 총 32문항으로 구성되어 있다. Cronbach α 계수를 바탕으로 신뢰도를 분석한 결과, 전체는 .92이며, 하위요인별로는 자기조절에서 .86, 사회성에서 .82, 도덕성에서 .86, 소명의식에서 .85, 학생에 대한 열정에서 .93, 그리고 교육적 신념에서 .87로 높게 나타났다.

[표 1] 교직 적성·인성검사에서 보고하는 Cronbach α
[Table 1] Cronbach α reporting in teachers' teaching aptitude and personality tests

대상	연구자	영역 (하위영역 수/문항 수)	신뢰도
예비 유아 교사	조운주 (2014)	1:계획성, 성실·책임감, 2:소명감·교직관, 3:언어·의사소통력, 지도성, 문제해결력 및 판단력, 4:심리적 안정성, 5:봉사·희생·협동성, 6:지식·정보력, 7:독립성·자주성, 8:사려성·타인존중, 9:창의·응용력	영역: .73-.93 전체:.96
예비 초등 교사	조주연 외 (2004)	1:교수능력(언어표현, 응용력, 판단력), 2:생활지도(신뢰감, 포용심, 지도력), 3:연구능력(관찰력, 탐구력, 문제해결), 4: 창의성(독창성, 융통성, 통찰력), 5: 소명감(소명감, 성실성, 열성), 6:도덕성(인간존중, 공정성, 도덕성)(18/180), 타당도포함(19/225)	영역: .63-.87 전체: 제시하지않음
	홍기칠 (2006)	1:교수능력(언어표현, 응용력, 판단력), 2:생활지도(신뢰감, 포용심, 지도력), 3:연구능력(관찰력, 탐구력, 문제해결), 4: 창의성(독창성, 융통성, 통찰력), 5: 소명감(소명감, 성실성, 열성), 6:도덕성(인간존중, 공정성, 도덕성)(18/180), 타당도포함(19/200)	신입학 영역: .63-.88 4학년 영역: .61-.89 전체: 제시하지않음
예비 교사	김정환 외 (2012) 예비 교직 적성·인성 검사	1:문제해결력·탐구력, 2:판단력, 3:독립성·자주성, 4:창의·응용력, 5: 심리적 안정성, 6:언어·의사소통능력, 7:지도성, 8: 공감·포용력, 9:지식·정보능력, 10:봉사·희생·협동성, 11:계획성, 12: 성실성·책임감, 13:소명감·교직관, 14:열정(14/350)	영역: .69-.86 전체: 제시하지않음
	김경령, 서은희 (2014) 인성 검사	1:보편적 인성(자기조절, 사회성, 도덕성), 2:교직 인성(소명의식, 학생에 대한 열정, 교육적 신념)(6/32)	영역: .82-.93 전체:.92
	서경혜 외	1:내적 인성(자기 조절, 반성적 실천, 지속적 배움), 2:사회적	영역: .61-.85

대상	연구자	영역 (하위영역 수/문항 수)	신뢰도
	(2013) 인성 검사	인성(존중, 의사소통, 협력), 3: 공동체적 인성(윤리의식, 정의감, 책임감)(9/26)	전체:
중등 예비 교사	김민웅, 김태훈 (2017) 인성 검사	1:교육적 열정, 2:교육적 신념, 3:소명의식, 4:자기조절, 5:자기개발, 6:도덕성, 7:책임감, 8:사회성, 9:소통(9/56)	영역: .73-.92 전체:.95

서경혜 외(2013)는 예비교사들의 교직 인성을 진단할 수 있는 총 26개의 문항으로 구성된 평가도구를 개발한 후, 사범대학, 일반대학 교직과정, 교육대학에 재학 중인 학생을 대상으로 평가도구의 신뢰도 및 타당도를 검증하였다. 신뢰도 검증을 위해 Cronbach α 계수를 산출한 결과, 대영역인 내적 인성, 사회적 인성, 그리고 공동체적 인성은 각각 .71, .81, 그리고 .85로 양호한 것으로 나타났다. 구체적으로 내적 인성의 구성요소인 자기조절은 .70, 반성적 실천은 .62, 지속적 배움은 .61로, 사회적 인성의 구성요소인 존중은 .66, 의사소통은 .64, 협력은 .62로, 그리고 공동체적 인성의 구성요소인 윤리의식은 .76, 정의감은 .78, 책임감은 .82로 나타났다.

또한 특성화고 예비교사를 대상으로 김민웅과 김태훈(2017)은 교직 인성 요소를 측정할 수 있는 9개 요인의 총 56문항으로 구성된 평가 도구를 개발 및 타당화하였다. 이 평가도구의 신뢰도를 검증하기 위해 Cronbach α 계수를 확인한 결과, 전체에서는 .95, 교육적 열정에서는 .92, 교육적 신념에서는 .88, 소명의식에서는 .90, 자기조절에서는 .83, 자기개발에서는 .84, 도덕성에서는 .75, 책임감에서는 .73, 사회성에서는 .74, 그리고 소통에서는 .84로 높게 나타났다.

[표 1]은 위의 내용을 요약한 것이며, 제시된 바와 같이 교직 적성·인성 검사에서 신뢰도 검증은 대부분 대학에 재학 중인 학생들을 대상으로 단일 오차요인만 고려함과 동시에 규준참조검사에 적합한 Cronbach α 계수만을 보고함으로써 검사의 전체 신뢰도를 과대추정할 수 있다는 문제가 제기된다(Brennan, 2001; Shavelson & Webb, 1991). 따라서 본 연구에서는 이에 대한 해결 방안의 하나로 교육대학원에 재학 중인 수학교육 전공의 원생들을 대상으로 다중 오차요인을 고려하는 일반화가능도 이론을 적용함으로써, 준거참조검사에 적합한 신뢰도 산출 방법과 다양한 신뢰도 계수가 동시에 적정 수준에 도달할 수 있는 효율적인 측정 조건을 제시하고자 한다.

2. 분류 결정에 대한 다양한 신뢰도

대부분의 검사기반 결정 절차들은 결정 규칙을 규정하는 한 개 이상의 분할점수를 수반하며, 분할점수의 선택은 결정 규칙을 정의하는데 가장 중요한 문제로 인식된다(Cizek, 1996). 분할점수를 설정하기 위해서는 다양한 방법이 가능하지만 미리 정해져 있는 비율만큼 피험자를 합격 또는 불합격으로 판정하는 규준적 준거설정방법은 본 연구에서 활용하고 있는 교직 적성·인성검사와 같은 자격시험에는 적절하지 못하다. 왜냐하면 규준참조점수는 다른 사람보다 잘했느냐 못했느냐로 해석되는 것이지, 어떤 특정 피험자가 어떤 역량 수준에 있는가로 해석될 수 없기 때문이다(Sireci, 2005). 또한 검사점수와 같은 연속점수가 하나 또는 둘 이상의 불연속점수인 분할점수에 대하여 해석되는 준거참조검사에서 검사 문항의 목적은 각각의 구성개념을 다르게 하여 전체 목표에 대한 달성여부를 판정하기 때문이다. 따라서 검사가 동일한 구성개념을 측정하는지에 적합한 신뢰도인 Cronbach α 계수는 적절하지 않다. 즉, 이러한 신뢰도 계수는 피험자들의 순위를 매기거나 순서를 정해야 하는 경우의 일관성을 보여주기 때문에 규준참조점수 해석에서만 유용하다. 반면에 준거참조검사에서는 검사점수의 일관성에 초점을 두는 것이 아니라 피험자들의 합격-불합격 여부에 대한 결정이 얼마나 일관성이 있는가가 관심이 된다(Popham & Husek, 1969). 따라서 본 절에서는 분할점수에 대해 피험자를 분류하는 준거지향검사의 분류 일관성 및 정확도를 특징지을 수 있는 다양한 신뢰도 계수들을 살펴본다.

먼저 분류일관성 계수는 분류 결정을 하는데 사용되는 검사의 신뢰도로 피험자가 동일한 측정 과정을 통해 반복하여 검사를 치렀을 때, 같은 결과로 분류된 피험자의 비율을 나타낸다(Livingston & Lewis, 1995). Swaminathan 외(1974)는 일관성 계수로 우연에 의한 오차를 제외한 Cohen(1960)의 카파(κ)를 대안으로 제시하였으며, 식(1)과 같다.

$$\kappa = \frac{P_A - P_c}{1 - P_c} \quad (1)$$

여기서 P_A는 두 번의 시행에서 같은 결과인 불합격-불합격과 합격-합격으로 분류된 피험자의 관찰 비율이며, P_c는 우연에 의해 분류 결정이 일치할 확률을 나타낸다. 그러나 검사를 두 번 시행하는 경우는 현실적으로 어려움이 많기 때문에, 한 번의 시행으로 κ와 P_A를 추정하기 위해 측정모형에 기반한 관찰점수의 분포에 대한 가정을 통해 일관성 계수를 산출하는 여러 가지 방법이 제시되었다.

Subkoviak(1976)은 같은 배점의 이항 선택형 문항으로 구성된 검사에서 사용할 수 있는 계수를, Huynh(1976)은 진점수를 베타분포로, 오차점수는 이항분포로 가정하여 일치도 계수를, Hanson과 Brennan(1990)은 진점수를 4모수 베타분포로, 오차점수는 이항분포로 가정하여 이분 문항에서의 분류 일치도 계수를 산출하였다. 반면에 Livingston과 Lewis(1995)는 이분문항에만 적용되었던 베타이항모형에 효과적인 검사 길이의 개념을 도입하여 이항분포에 적용함으로써, 배점이 서로 다른 이분 문항이나 다분 문항으로 구성된 검사에 적용할 수 있는 분류 일치도 계수를 산출하였다.

다음으로 분류정확도 계수는 분할점수로부터 피험자들이 떨어진 정도를 의미하며, 관찰점수와 진점수 상에서의 분할점수를 이용한 합격여부에 대한 분류 간 일치 정도를 나타낸다(Livingston & Lewis, 1995). Livingston(1972)은 고전검사이론의 신뢰도 계수를 일반화시켜 식(2)와 같은 정확도 계수를 산출하였다.

$$\kappa^2 = \frac{\sigma_T^2 + (\mu - C)}{\sigma_X^2 + (\mu - C)} \quad (2)$$

여기서 σ_T^2는 피험자의 진점수 분산, σ_X^2는 피험자의 관찰점수 분산, μ는 검사집단의 평균, 그리고 C는 분할점수를 나타낸다.

Brennan과 Kane(1977)은 일반화가능도 이론의 의존도 계수 ϕ와 분할점수에 따른 의존도 계수 $\phi(\lambda)$를 준거참조검사의 신뢰도로 제안하였으며, 더 일반적인 계수인 $\Phi(\lambda)$는 식 (3)에 의해 얻을 수 있다.

$$\phi(\lambda) = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\triangle)} \quad (3)$$

여기서 $\sigma^2(p)$는 전집점수 분산, μ는 평균점수, λ는 분할점수, 그리고 $\sigma^2(\triangle)$는 측정대상의 관찰점수와 전집점수 차이에 대한 분산으로 전집점수 분산을 제외한 모든 분산을 포함한다. 특히 이분 문항으로 점수가 부여되는 경우에 의존도 계수는 $\lambda = \overline{X}$일 때, 식(3)은 KR-21과 일치하게 된다. 또한 식(3)에서 $\lambda = \mu$인 경우의 $\phi(\lambda)$를 ϕ로 정의하며, 식(4)에 의해 얻을 수 있다.

$$\phi = \frac{E_p(\mu_p - \mu)^2}{E_I E_p(X_{pI} - \mu)^2} = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\triangle)} \quad (4)$$

여기서 E_p와 $E_I E_p$는 각각 문항 및 문항과 피험자에 대한 기댓값을 의미한다.

신뢰도 계수로 산출되는 ϕ와 $\phi(\lambda)$는 연구자의 관심에 따라 구분지어질 수 있다. Brennan(1984)에 따르면 $\phi(\lambda)$는 준거참조검사의 절차에 근거한 분류의 신뢰도 추정이며, Φ는 검사절차가 그러한 분류의 신뢰도에 기여한 정도를 추정한다. 즉, $\phi(\lambda)$는 분할점수의 개수만큼의 계수를 산출함과 동시에, 가장 적정한 계수를 산출할 수 있는 측정 조건과 분할점수의 조합을 제시하는데 이용할 수 있다(김성숙, 김양분, 2001). 이상에서 살펴본 바와 같이 일반화가능도 이론을 바탕으로 한 의존도 계수는 한 번의 검사를 통해 얻은 결과만으로도 비교적 용이한 계산으로 신뢰도 계수를 산출함을 알 수 있다.

Ⅲ. 연구방법

1. 분석 자료

본 연구는 교직 적성·인성 검사 실시가 의무화된 2013년부터 최근 2017년 5월까지 수도권에 소재한 I교

육대학원에서 수학교육을 전공한 33명의 원생을 대상으로 실시한 교직 적성·인성 검사 결과를 분석 자료로 활용하였다. 교직 적성·인성 검사는 학기에 따라 5월 또는 11월의 마지막 주 목요일에 40분간 지필검사로 실시되었다. 또한 채점은 교직 적성·인성 검사 표준안(김정환 외, 2012)에서 제시한 절차에 따라 답변일관성 문항들에 대해서는 0점부터 4점까지로 재코딩 하였으며, 교직 적성·인성 검사를 두 번 치른 원생의 경우는 첫 번째 결과만을 분석 자료에 포함시켰다. 연도별로 검사에 응답한 수학교육 원생들의 수 및 비율은 [표 2]와 같다. 또한 성별 구성은 남자 원생의 경우 14명으로 약 42%, 그리고 여자 원생의 경우 19명으로 약 58%였다.

[표 2] 연도별 교육대학원생 수 및 비율
[Table 2] Item scores and scoring domains of a mathematical creativity test

년도-학기	빈도	퍼센트	년도-학기	빈도	퍼센트
2013-1	10	30	2015-2	4	12
2013-2	9	27	2016-1	3	9
2014-1	1	3	2016-2	1	3
2014-2	1	3	2017-1	3	9
2015-1	1	3	합계	33	100

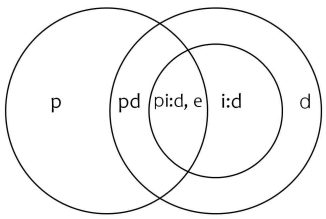
2. 분석 방법

일반화가능도이론은 측정 오차에 영향을 주는 다중 오차요인을 찾고, 설계모형에 따라 각 효과의 분산성분을 추정하는 일반화 연구(G-연구)와 G-연구 결과를 토대로 오차를 최소화함으로써 측정 모형을 최적화하는 결정 연구(D-연구)로 나눈다. 구체적으로 G-연구는 허용가능한 관찰전집(universe of admissible observation)과 관계된 잠재적 오차의 근원을 밝히고 오차의 양을 구하기 위해 수행한다. 여기서 허용가능한 관찰전집이란 연구에서 고려하는 요인 또는 변인과 조건의 범위를 의미한다. 반면에 D-연구에서는 일반화 전집(universe of generalization)을 규정하고 측정의 변동에 영향을 주는 관측 조건을 조정하거나 결정하기 위해 수행한다. 일반화 전집이란 의사결정자가 D-연구를 수행하기 위해 일반화하고자 하는 전집을 의미한다(김성숙, 김양분, 2001).

1) G-연구 설계

본 연구에서는 모든 수학교육 원생(p)에게 14개의 적성·인성 하위 영역(d) 별로 15개의 문항(i)이 포함되어 있는 총 210문항의 동일한 교직 적성·인성 검사가 시행되었으므로, 이는 일반화가능도 이론의 G-연구 설계인 2국면 내재모형(two-facet nested design) $p \times (i : d)$로 나타낼 수 있다. 여기서 원생은 측정의 대상으로 허용 가능한 변동요인을 나타내는 국면의 수에는 포함시키지 않으며, 오차 국면에는 영역과 문항이 포함된다. 또한 일반화가능도 이론에서 전집(universe)은 표집 단위를 측정 대상으로 하는 모집단과 다르게 측정 대상의 측정 조건들에 대한 일반화과정을 포함한다. 즉, 원생은 무한한 모집단에서 무작위로 추출되었으며, 영역과 문항은 무한한 영역과 문항 전집에서 표집되었다고 간주할 수 있으므로, 모두 무선효과(random effects)로 가정하였다. 따라서 어떤 원생이 14개 하위 영역의 각 문항에 응답한 점수는 식(1)과 같이 차례대로 총 평균, 원생 효과, 영역 효과, 각 영역 내 문항 효과, 영역과 원생의 상호작용 효과, 그리고 잔차 효과로 분해할 수 있으며, 각 효과를 도식화하면 [그림 1]과 같다.

$$\begin{aligned} X_{pid} = {} & \mu \\ & + (\mu_p - \mu) \\ & + (\mu_d - \mu) \\ & + (\mu_{id} - \mu_d) \\ & + (\mu_{pd} - \mu_p - \mu_d + \mu) \\ & + (\mu_{pid} - \mu_{pd} - \mu_{id} + \mu_p) \end{aligned} \quad (1)$$



[그림 1] $p \times (i : d)$의 G-연구 설계
[Fig. 1] G-study for the $p \times (i : d)$ design

또한 어떤 원생의 교직 적성·인성 검사점수의 분산은 식 (2)와 같이 5개의 독립적인 분산성분의 합으로

설명할 수 있다. 여기서 σ_p^2는 교직 적성·인성 검사점수에서 나타난 원생 간의 차이, σ_d^2 는 영역 간 차이, $\sigma_{i:d}^2$는 각 영역 내에서의 문항 간 차이, σ_{pd}^2는 영역에 따라 원생이 다르게 응답한 정도, 그리고 $\sigma_{pi:d,e}^2$는 영역 내에서 원생과 문항의 상호작용과 잔차에 의한 변동을 나타낸다.

$$\sigma_x^2 = \sigma_p^2 + \sigma_d^2 + \sigma_{i:d}^2 + \sigma_{pd}^2 + \sigma_{pi:d,e}^2 \quad (2)$$

각 분산성분 추정치는 평균제곱의 기댓값을 사용하여 식(3)과 같이 얻을 수 있다. 여기서 MS는 평균제곱, n_p, n_d, n_i는 각각 피험자, 영역, 그리고 문항 수를 나타낸다.

$$\begin{aligned}
\sigma_p^2 &= (MS_p - MS_{pd})/n_i n_d \quad (3)\\
\sigma_d^2 &= (MS_d - MS_{i:d} - MS_{pd} + MS_{pi:d})/n_p n_i\\
\sigma_{i:d}^2 &= (MS_{i:d} - MS_{pi:d})/n_p\\
\sigma_{pd}^2 &= (MS_{pd} - MS_{pi:d})/n_i\\
\sigma_{pi:d}^2 &= MS_{pi:d}
\end{aligned}$$

이렇게 추정된 분산성분 추정치는 G-연구의 최종 결과이며, 이를 바탕으로 다양한 D-연구를 수행할 수 있다.

2) D-연구 설계

본 연구에서는 D-연구 설계를 G-연구 설계와 동일하게 적용하였다. 다만 D-연구에서의 점수는 측정 대상인 원생의 모든 관찰점수에 대한 평균을 사용하므로, $p \times (I:D)$와 같이 오차 국면을 대문자로 표현한다. 또한 Brennan(2001)에서 제시한 표기법에 따라 G-연구에서 문항과 영역의 수가 각각 n_i와 n_d이었다면, D-연구에서 사용하는 문항과 영역의 수는 각각 n_i'와 n_d'로 표현한다.

$p \times (I:D)$ 설계에서의 분산성분은 식(4)에 의해 얻을 수 있다. 여기서 $\overline{\alpha}$는 G-연구 설계에서의 분산성분과 같으나 D-연구 설계의 분산성분임을 강조하기 위해 표기한 것이며, $d(\overline{\alpha})$는 $\overline{\alpha}$가 p인 경우에는 1로, 나머지 경우는 p를 제외한 D-연구 설계에서 오차 국면의 표본 수를 나타낸다. 구체적으로 절대오차분산($\sigma^2(\triangle)$)은 식(5)에서, 그리고 의존도 계수(Φ)는 식 (6)에 의해 얻을 수 있다.

$$\sigma_{\overline{\alpha}}^2 = \frac{\sigma_{\overline{\alpha}}^2}{d(\overline{\alpha})}, \quad (4)$$

$$\sigma^2(\triangle) = \frac{\sigma_d^2}{n_d'} + \frac{\sigma_{i:d}^2}{n_i' n_d'} + \frac{\sigma_{pd}^2}{n_d'} + \frac{\sigma_{pi:d,e}^2}{n_i' n_d'}, \quad (5)$$

$$\begin{aligned}
\phi &= \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2(\triangle)} \quad (6)\\
&= \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_d^2}{n_d'} + \frac{\sigma_{i:d}^2}{n_i' n_d'} + \frac{\sigma_{pd}^2}{n_d'} + \frac{\sigma_{pi:d,e}^2}{n_i' n_d'}}.
\end{aligned}$$

또한 분할점수 λ를 기준으로 한 의존도 계수($\phi(\lambda)$)는 구체적으로 식(7)에 의해 얻을 수 있다. 여기서 μ는 교직 적성·인성 검사의 평균점수로 본 연구에서는 666점으로 산출되었으며, 분할점수 λ는 현재 이 검사를 사용하고 있는 학교 현장의 최저 분할점수를 반영하여 총점의 60%인 504점부터 80%인 716점까지 다양하게 변화시켰다.

$$\begin{aligned}
\phi(\lambda) &= \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma^2(\triangle)} \quad (7)\\
&= \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \frac{\sigma_d^2}{n_d'} + \frac{\sigma_{i:d}^2}{n_i' n_d'} + \frac{\sigma_{pd}^2}{n_d'} + \frac{\sigma_{pi:d,e}^2}{n_i' n_d'}}.
\end{aligned}$$

Brennan 외(1995), Lee와 Kantor(2007), Norcini(1999), Schoonheim-Klein 외(2008), Xi(2007), 그리고 Yin과 Sconing(2008)은 의존도 계수를 해석할 때, 측정치의 정확성을 나타내는 표준오차(Standard Error of Measurement, SEM)도 신뢰도 계수의 일종으로 함께 고려할 것을 제안하였다. SEM은 식(6)에 제시되어 있는 절대오차분산($\sigma^2(\triangle)$)에 근호를 취함으로써 얻을 수 있다.

또한 교직 적성·인성 검사점수의 재현성이 아니라

적격과 부적격 결정에 대한 재현성(reproducibility)을 나타내는 신뢰도 계수로 수정된 의존도 계수(adjusted dependability coefficients, ϕ_{adj})를 사용하였다. ϕ_{adj}는 식(6)에서 평균으로부터의 편차 제곱을 분할점수로부터의 편차 제곱으로 대체함으로써 얻을 수 있다(Govaerts et al., 2002).

또한 적정 수준의 신뢰도에 대해 Brennan(2001)과 Fyans(1983)는 의존도 계수의 경우 .70, Norcini(1999)는 검사 종류에 상관없이 .80으로, 그리고 영재 선발 검사의 경우에는 .90(Hopkins, 1997; Salvia et al., 2012)으로 연구자에 따라 다양하게 제시하고 있지만, 고부담 검사(high risk test)일수록 신뢰도 계수가 높아야 한다는 점에서는 연구자들의 견해는 일치하고 있다(Nunnally & Bernstein, 1994). 따라서 본 연구에서는 적정 수준의 신뢰도를 원자료의 의존도 계수인 .872로 정하였다. 또한 SEM은 교직 적성·인성 검사의 문항이 4점 척도이므로 적어도 1점 이내에서 원생들의 점수에 대한 신뢰로운 추론을 하기 위하여 0.24로 정하였다. G-연구와 D-연구 설계의 분산성분 추정치, 절대오차분산, 의존도 계수, 분할점수에 따른 의존도 계수, 측정의 표준오차, 그리고 수정된 의존도 계수는 GENOVA(GENeralized analysis Of VAriance System)* 프로그램과 엑셀의 매크로를 사용하여 산출하였다.

Ⅵ. 결과 분석 및 논의

1. G-연구 분석 결과

교직 적성·인성 검사 결과에서 원생의 점수에 영향을 주는 요인들, 즉 원생(p), 영역(d), 영역 내 문항($i:d$), 원생과 영역의 상호작용(pd), 그리고 영역 내 원생과 문항의 상호작용을 포함한 잔차($pi:d,e$)의 분산성분 크기와 각 효과의 영향력은 [표 3]과 같다. 영역 내 문항 효과와 영역 내 원생과 문항의 상호작용을 포함한 잔차 효과는 각각 33.97%와 54.21%로 상대적으로 크게 나타났다. 영역 내 문항 효과는 원생들이 영역 내 문항들에 대해 다르게 응답하고 있음을 의미하며, 이는 준거참조검사의 특성이 각 문항의 독립적 구성개념을 나타내기 때문에 문항 간 차이가 크다는 것을 뒷받침한다고 해석할 수 있다(김성숙, 김양분, 2001). 또한 영역 내 원생과 문항의 상호작용을 포함한 잔차 효과는 영역 내 문항에 대해 원생들이 다르게 응답하고 있음을 의미하며, 동시에 본 연구 설계에서 고려하지 못한 다른 오차 요인들(예: 시행 연도, 시행 시간, 시행 장소 등)의 필요성에 대한 간접적인 설명으로 해석할 수 있다.

[표 3] $p \times (i:d)$ 설계 분석 결과
[Table 3] Results for $p \times (i:d)$ design

효과	자유도	제곱합	평균제곱	분산추정치(%)
p	32	428.154	13.379	0.058(5.81)
d	13	236.217	18.171	0.012(1.20)
$i:d$	196	2300.755	11.739	0.339(33.97)
pd	416	525.339	1.263	0.048(4.81)
$pi:d,e$	6272	3395.777	0.541	0.541(54.21)
전체분산	6929	6886.242	45.093	0.998(100.00)

주.p는 원생, d는 영역, i는 문항, 그리고 e는 잔차를 나타냄.

반면에 영역 효과는 1.20%로 상대적으로 작게 나타났지만, 원생과 영역의 상호작용 효과는 4.81%로 영역 효과보다 4배 크게 나타났다. 이는 전체 원생의 인성검사 점수가 영역에 따라 큰 차이를 보이지는 않으나, 영역에 따라 원생들의 순위에는 변동이 있다고 해석할 수 있다. 즉, 한 영역에서 높은 점수를 보였던 원생이 다른 영역에서는 낮은 점수를 보이고 있다는 것이다. 또한 의존도 계수에 긍정적인 영향을 미치는 전집점수 분산의 크기를 나타내는 원생 효과는 5.81%로 나타났으며, 이를 통해 검사 점수에서 원생들의 인성 차이가 반영되어 있다고 해석할 수 있다.

2. D-연구 분석 결과

G-연구 분석 결과에서 산출된 [표 3]의 분산 추정치를 바탕으로 원 자료의 표본 수와 동일하게 수행한 D-연구 분석 결과, 전집분산은 0.058, 절대오차는 0.008 $(=0.012/14+0.339/(14\cdot 15)+0.048/14+0.541/(14\cdot 15))$, 그리고 의존도 계수는 .872$(=0.058/(0.058+0.008))$로 나타났다. 본 연구에서는 적정 수

* GENOVA 프로그램은 https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs에서 무료로 다운로드 받을 수 있다.

준의 신뢰도를 원자료의 의존도 계수인 .872로 정하고, 다양한 D-연구 분석을 위해 영역 수와 영역 내 문항 수를 각각 2개부터 원 자료의 수까지 증가시켰다. 김정환 외(2012)의 교직 적성·인성 검사 표준안은 총 11단계에 걸쳐 개발되는 동안 영역 수는 다양하게 변화되었지만, 영역 내 문항 수는 동일하게 구성하였으므로, 이를 반영하여 본 분석에서도 영역 내 문항 수를 동일하게 유지하였다. 또한 김정환 외(2012)의 교직 적성·인성 검사를 실시하고 있는 실제 학교 별 분할점수를 고려하여 평균점수 척도 상에서 이루어진 GENOVA 프로그램의 분석 결과에 대한 해석의 편의를 위하여 분할점수는 총점으로 표시하였다. 즉, 총점 840점을 기준으로 60%에 해당하는 504점부터 80%에 해당하는 716점까지 1점씩 분할점수를 증가시키면서 의존도 계수를 산출하였다. 이 중 [표 4]는 지면 제약 상 504점에서 의존도 계수가 .872이상인 경우 중 일부 결과를 제시하였다.

[표 4] 다양한 $p \times (I:D)$ 설계 분석 결과
[Table 4] Various results for $p \times (I:D)$ design

영역수2:	문항수:8	9	10	11	13	15
$\widehat{\sigma^2}(p)$	0.058	0.058	0.058	0.058	0.058	0.058
$\widehat{\sigma^2}(\triangle)$	0.085	0.079	0.074	0.070	0.064	0.059
$\hat{\phi}$	0.405	0.423	0.439	0.452	0.475	0.494
$\hat{\phi}(\overline{X})$	0.243	0.274	0.300	0.323	0.360	0.389
$\hat{\phi}(504)$	*0.880*	*0.888*	*0.894*	*0.900*	*0.908*	*0.914*
$\hat{\phi}(588)$	0.659	0.679	0.695	0.709	0.731	0.747
$\hat{\phi}(630)$	0.398	0.427	0.451	0.471	0.504	0.529
$\hat{\phi}(672)$	0.249	0.280	0.306	0.329	0.365	0.394
$\hat{\phi}(716)$	0.499	0.525	0.547	0.565	0.594	0.616
영역수6:	문항수:2	5	8	10	13	15
$\widehat{\sigma^2}(p)$	0.058	0.058	0.058	0.058	0.058	0.058
$\widehat{\sigma^2}(\triangle)$	0.083	0.039	0.028	0.023	0.021	0.020
$\hat{\phi}$	0.409	0.595	0.671	0.713	0.731	0.745
$\hat{\phi}(\overline{X})$	0.225	0.516	0.621	0.675	0.699	0.716
$\hat{\phi}(504)$	*0.882*	*0.942*	*0.958*	*0.965*	*0.968*	*0.970*
$\hat{\phi}(588)$	0.748	0.871	*0.905*	*0.921*	*0.928*	*0.933*
$\hat{\phi}(630)$	0.393	0.646	0.729	0.771	0.788	0.802
$\hat{\phi}(672)$	0.230	0.521	0.625	0.679	0.702	0.719
$\hat{\phi}(716)$	0.495	0.716	0.786	0.810	0.834	0.845
영역수10:	문항수:2	5	8	10	13	15
$\widehat{\sigma^2}(p)$	0.058	0.058	0.058	0.058	0.058	0.058
$\widehat{\sigma^2}(\triangle)$	0.050	0.024	0.017	0.015	0.013	0.012
$\hat{\phi}$	0.536	0.710	0.773	0.796	0.819	0.830
$\hat{\phi}(\overline{X})$	0.424	0.668	0.747	0.776	0.803	0.815
$\hat{\phi}(504)$	*0.926*	*0.941*	*0.957*	*0.963*	*0.968*	*0.970*
$\hat{\phi}(588)$	0.839	*0.920*	*0.942*	*0.949*	*0.956*	*0.959*
$\hat{\phi}(630)$	0.572	0.766	0.825	0.846	0.865	*0.874*
$\hat{\phi}(672)$	0.429	0.672	0.750	0.778	0.805	0.818
$\hat{\phi}(716)$	0.653	0.817	0.863	*0.880*	*0.896*	*0.903*
영역수14	문항수:2	5	8	10	13	15
$\widehat{\sigma^2}(p)$	0.058	0.058	0.058	0.058	0.058	0.058
$\widehat{\sigma^2}(\triangle)$	0.036	0.017	0.012	0.011	0.009	0.008
$\hat{\phi}$	0.618	0.774	0.826	0.845	0.864	*0.872*
$\hat{\phi}(\overline{X})$	0.542	0.748	0.810	0.833	0.848	0.863
$\hat{\phi}(504)$	*0.947*	*0.975*	*0.982*	*0.984*	*0.986*	*0.987*
$\hat{\phi}(588)$	*0.882*	*0.942*	*0.958*	*0.963*	*0.968*	*0.970*
$\hat{\phi}(630)$	0.669	*0.826*	0.871	*0.886*	*0.901*	*0.908*
$\hat{\phi}(672)$	0.547	0.751	0.813	0.835	0.855	0.865
$\hat{\phi}(716)$	0.736	0.864	*0.900*	*0.912*	*0.924*	*0.929*

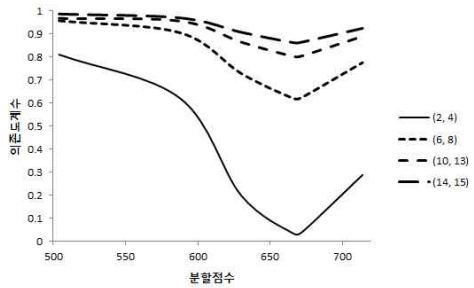
주1. $\hat{\phi}(\overline{X})$는 검사의 평균점수인 666점을 분할점수로 했을 때의 의존도 계수를 나타냄.
주2. 기울임체는 의존도 계수가 적정 수준의 신뢰도인 .872에 도달한 경우를 나타냄.

[표 4]에 제시된 바와 같이 영역 수를 6개로 제한하는 경우에는 현 교직 적성·인성 검사의 문항 수와 동일하게 영역별 문항 수를 15개로 정하면 총 문항 수는 90개이고 이 때 의존도 계수는 .745로 나타났다. 반면에 영역 수가 10개인 경우에는 각 영역별로 문항 수를 8개로 감소시켜 총 문항 수가 80개가 되지만, 이 때 의존도 계수는 .773으로 총 문항 수가 90개인 경우보다 높게 나타났다.

영역 수가 2개이며 분할점수를 504점으로 할 때는 문항 수가 8개 이상인 경우에 의존도 계수가 적정 수준인 .872이상인 것으로 나타났지만, 분할점수를 588점 이상으로 하는 경우에는 15개 문항에서도 의존도 계수가 적정 수준에 도달하지 못하는 것으로 나타났다. 반면에 6개 영역에서는 분할점수가 504점인 경우에 2문항 이상, 분할점수가 588점인 경우에는 8문항 이상인 경우에

의존도 계수가 .872에 도달하였으며, 분할점수가 630점 이상인 경우에는 의존도 계수가 적정 수준에 도달하지 못하는 것으로 나타났다. 또한 영역의 수와 문항 수에 상관없이 분할점수가 평균과 비슷한 672점인 경우에 의존도 계수는 .872에 도달하지 못하는 것으로 나타났다.

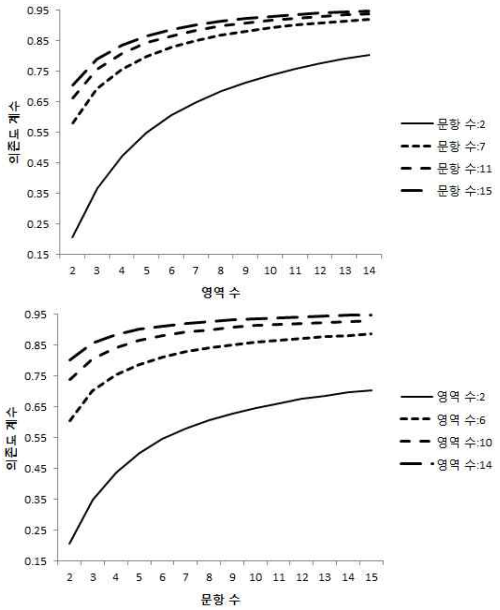
또한 영역 수와 상관없이 분할점수를 평균점수로 할 때는 의존도 계수가 적정 수준에 도달하지 못하는 것으로 나타났으며, 영역 수를 10개와 14개로 정하면 각각 문항 수가 10개 이상과 8개 이상일 때 적정 수준에 도달하는 것으로 나타났다. 반면에 영역 수가 14개인 경우에 분할점수를 504점과 588점으로, 영역 수가 10개인 경우에 분할점수를 504점으로 정하면 모든 문항 수의 경우에 대해, 그리고 영역 수가 10개인 경우에 분할점수를 588점으로 정하면 문항 수가 5개 이상일 때 의존도 계수는 적정 수준에 도달하는 것으로 나타났다.



[그림 1] 분할점수에 따른 의존도 계수 변화
[Fig. 1] Changes of $\hat{\phi}(\lambda)$ depending on various cut scores

[표 4]에서 분할점수 변화에 따른 의존도 계수의 일부를 제시하면 [그림 1]과 같다. [그림 1]에서 각 괄호 안의 숫자는 순서대로 영역 수와 문항 수를 나타낸다. 이를 통해 총 문항 수가 많을수록 의존도 계수는 높게 산출되며, 분할점수가 검사의 평균점수와 같은 경우에 의존도 계수는 가장 낮게 산출되며, 분할점수와 평균점수와의 차이가 클수록 의존도 계수는 높게 산출됨을 알 수 있다.

[그림 2]는 김정환 외(2012)가 분할점수로 제시한 598점에서 [표 4]에 제시한 영역 수와 문항 수를 반영한 모든 의존도 계수를 제시하였다. 영역 수를 5개 이하로 제한하는 경우, 현 교직 적성·인성검사와 동일하게 영역별 문항 수를 15개로 하는 경우에도 적정 수준의 신뢰도인 .872에 도달하지 못하는 것으로 나타났다.



[그림 2] 문항 수와 영역 수에 따른 의존도 계수 변화
[Fig. 2] Changes of $\hat{\phi}$ depending on various numbers of both items and domains

반면에 영역 수를 6개, 7개, 8개, 9개, 10개, 11개, 12개, 13개, 14개로 정하는 경우에는 영역별 문항 수가 각각 12개, 10개, 8개, 7개, 6개, 5개, 5개, 4개, 4개일 때 적정 수준의 신뢰도인 .872에 도달하는 것으로 나타났다. 그러나 영역 수를 고정할 때는 문항 수가, 문항 수를 고정할 때는 영역 수가 증가할수록 절대오차분산의 크기는 작아지고 의존도 계수는 커지는 것으로 나타났다.

[표 5]는 교직 적성·인성 검사점수를 504점부터 716점까지 1점씩 분할점수를 증가시키면서 SEM과 수정된 의존도 계수를 산출한 결과 중, 지면 제약 상 일부를 제시하였다. 문항 수와 상관없이 영역 수를 2개 또는 6개로 정할 때는 어떤 분할점수에서도 SEM이 .24이하면서 동시에 수정된 의존도 계수가 .872에 도달하는 경우는 없는 것으로 나타났다. 반면에 김정환 외(2012)가 제

시한 분할점수 598점에서는 영역 수 10개이며 문항 수 10개 이상인 경우에, 영역 수 12개이며 문항 수 8개 이상인 경우에, 그리고 영역 수 14개이며 문항 수 6개 이상인 경우에 수정된 의존도 계수는 .872에 도달하는 것으로 나타났다. 특히 총 문항 수가 96개인 영역 수 12개이며 문항 수 8개인 경우의 의존도 계수는 .880으로 총 문항 수 100개인 각각 영역 수와 문항 수가 10개인 경우의 의존도 계수인 .876보다 높게 나타났다. 또한 원자료와 같은 14개 영역에 영역별 문항 수가 15개인 경우에 수정된 의존도 계수가 적정 수준에 도달하는 경우로는 평균점수보다 낮을 때는 분할점수 615점 이하이며, 평균점수보다 높을 때는 분할점수 716점 이상인 것으로 나타났다. 또한 수정된 의존도 계수는 분할점수의 변화에 따라 [그림 1]의 의존도 계수와 유사한 패턴을 나타냈다.

[표 5] 분할점수에 따른 $p \times (I : D)$ 설계의 수정된 의존도 계수
[Table 5] SEM and Adjusted dependability coefficients associated with cut scores for $p \times (I : D)$ design

영역수6: 문항수:2	4	5	6	7	8	10	12	13	14	15
SEM	0.288	0.216	0.197	0.185	0.176	0.167	0.152	0.149	0.146	0.143	0.141
$\widehat{\phi_{adj}}(598)$	0.555	0.690	0.726	0.751	0.771	0.786	0.808	0.824	0.830	0.836	0.840
$\widehat{\phi_{adj}}(615)$	0.412	0.556	0.597	0.629	0.653	0.673	0.703	0.724	0.733	0.741	0.747
$\widehat{\phi_{adj}}(716)$	0.408	0.552	0.593	0.625	0.650	0.670	0.700	0.721	0.730	0.737	0.744
영역수10: 문항수:2	4	5	6	7	8	10	12	13	14	15
SEM	0.224	0.167	0.154	0.144	0.136	0.130	0.122	0.115	0.113	0.111	0.110
$\widehat{\phi_{adj}}(598)$	0.675	0.788	0.815	0.834	0.849	0.860	*0.876*	*0.887*	*0.891*	*0.895*	*0.898*
$\widehat{\phi_{adj}}(615)$	0.538	0.676	0.712	0.738	0.759	0.774	0.798	0.814	0.821	0.826	0.831
$\widehat{\phi_{adj}}(716)$	0.534	0.672	0.709	0.735	0.756	0.772	0.795	0.812	0.818	0.824	0.829
영역수12: 문항수:2	4	5	6	7	8	10	12	13	14	15
SEM	0.204	0.153	0.140	0.131	0.124	0.119	0.111	0.105	0.103	0.101	0.099
$\widehat{\phi_{adj}}(598)$	0.714	0.817	0.841	0.858	0.871	*0.880*	*0.894*	*0.904*	*0.907*	*0.911*	*0.913*
$\widehat{\phi_{adj}}(615)$	0.583	0.714	0.748	0.772	0.790	0.805	0.826	0.840	0.846	0.851	0.855
$\widehat{\phi_{adj}}(716)$	0.579	0.711	0.745	0.769	0.788	0.802	0.823	0.838	0.844	0.849	0.853
영역수14 문항수:2	4	5	6	7	8	10	12	13	14	15
SEM	0.190	0.141	0.130	0.121	0.115	0.110	0.105	0.097	0.095	0.094	0.089
$\widehat{\phi_{adj}}(598)$	0.744	0.839	0.861	*0.876*	*0.887*	*0.896*	*0.908*	*0.916*	*0.920*	*0.922*	*0.925*
$\widehat{\phi_{adj}}(615)$	0.620	0.745	0.776	0.798	0.815	0.828	0.847	0.860	0.865	0.869	*0.873*
$\widehat{\phi_{adj}}(716)$	0.616	0.742	0.773	0.796	0.812	0.826	0.845	0.858	0.863	0.868	*0.872*

주. 기울임체는 SEM이 0.24 이하이며, 수정된 의존도 계수가 적정 수준의 신뢰도인 0.872에 도달한 경우를 나타냄.

V. 결론 및 제언

본 연구는 교직 적성·인성 검사 표준안(김정환 외, 2012)을 수학교육 전공 원생들에게 실시한 결과에 일반화가능도 이론을 적용하여 검사점수에 영향을 미치는 오차 요인들의 상대적인 영향력을 분석하고, 분할점수 설정에 따라 다양한 신뢰도 계수들이 동시에 적정 수준인 .872에 도달할 수 있는 효율적인 측정 조건을 탐색하였다. 이를 토대로 교직 적성·인성 검사의 일반화가능도와 본 연구에서 적용한 분석 방법의 교육현장 적용

가능성에 대해 논하면 다음과 같다.

첫째, 교직 적성·인성 검사에 영향을 미치는 요인들의 상대적인 영향력을 탐색한 결과, 그 크기는 잔차, 영역 내 문항, 원생, 원생과 영역의 상호작용, 그리고 영역 순으로 나타났다. 잔차 분산이 전체 분산 중 가장 높은 비율을 차지하였다는 의미는 본 연구 설계에서 검사지 자체에서 발생할 수 있는 오차 이외의 다른 오차 요인들을 고려하지 않았기 때문이라고 해석할 수 있다. 따라서 교직 적성·인성 검사에서 보다 체계적으로 오차 분산을 감소시키기 위해서는 본 논문의 설계에서 고려하지 못한 다양한 오차 요인들, 예를 들면 시행 시기, 시행 횟수, 시행 시간 등이 설계에 반영되어야 한다. 이처럼 일반화가능도 이론을 적용하면 고전검사이론을 적용할 때는 변별할 수 없었던 다양한 오차 요인들을 파악할 수 있다. 이는 각 교원양성기관의 교직 적성·인성 검사 개발자들에게 시사점을 준다. 교직 적성·인성 검사를 개발할 때는 검사 결과에 영향을 미칠 수 있는 오차 요인들을 미리 밝혀내고, 측정하고자 하는 교직 적성·인성 이외의 다른 오차 요인들을 없애려는 노력을 해야 한다. 또한 실제 교육현장에서 얻게 되는 다양한 검사 점수들에 피험자의 능력 차이 외에 문항의 난이도처럼 하나의 오차 요인만 영향을 미치는 것이 아니라 다양한 오차 요인들이 영향을 미칠 수 있음을 고려한 설계를 바탕으로 해석해야 함을 제안할 수 있다.

둘째, 다양한 신뢰도 계수를 고려하여 교직 적성·인성 검사의 효율적인 측정 조건을 탐색한 결과, 분할점수 598점에서는 210문항으로 구성된 교직 적성·인성 검사를 12개 영역과 영역별 문항 수가 8개인 총 96개 문항으로 축소할 수 있는 것으로 나타났다. 또한 원자료와 같은 14개 영역과 영역별 문항수가 15개인 경우에는 분할점수를 평균점수보다 낮게는 615점으로, 그리고 높게는 716점으로 할 때 다양한 신뢰도 계수가 모두 적정 수준에 도달하는 것으로 나타났다. 지금까지 검사 개발 및 검사 결과 분석에서는 대부분 고전검사이론을 바탕으로 한 Cronbach α 계수가 보고되었으며, 준거참조검사에 일반화가능도 이론을 적용한 기존의 연구들에서도 대부분 의존도 계수만이 보고되었다. 그러나 본 연구에서는 검사 점수에 영향을 미치는 다양한 오차 요인, 추정치의 정확성, 검사점수가 아닌 적격과 부적격 판정에 대한 재현성을 나타내는 수정된 의존도 계수를 함께 고려하여 최적의 측정 조건을 탐색하였다는데 의의가 있다. 즉, 본 연구는 교육현장에서 검사의 효율성을 높일 수 있는 최적의 측정 조건을 탐색하는 방법으로 보다 정확하고 체계적인 분석을 시도하였다. 따라서 본 연구에서 적용한 방법론적 틀은 각 교원양성기관의 특성에 맞추어 교직 적성·인성 검사를 수정·보완하고자 할 때뿐만 아니라, 복합적인 자료나 복잡한 측정상황의 점수에 대한 설계와 분석을 수행하는 데에도 유용하게 활용될 수 있다.

셋째, 교직 적성·인성 검사 점수에서 다양한 신뢰도 계수는 분할점수 설정에 따라 많은 차이를 나타냈다. 또한 분할점수가 검사의 평균점수와 같은 경우에 신뢰도 계수는 가장 낮게 산출되었으며, 분할점수와 평균점수와의 차이가 클수록 신뢰도 계수는 높게 산출되었다. 따라서 준거참조검사를 사용하는 경우 평균점수를 분할점수로 설정하는 것은 측정학적 특성의 변화에 대한 경험적인 근거를 바탕으로 적정하지 않음을 밝혔다. 따라서 준거참조검사를 최소수행수준에 대한 달성을 목적으로 하는 경우와 선발 배치 등을 목적으로 하는 경우를 구분하여 평균점수보다 높게 또는 낮게 설정할 필요가 있다. 특히 현재 사용되는 교직 적성·인성 검사가 교직에 적합하지 않은 예비교사를 걸러내는 장치로서의 역할을 하기 위해서는 측정학적 특성의 변화에 대한 경험적인 근거를 바탕으로 분할점수를 평균점수보다 상향조정하는 것을 제안할 수 있다.

마지막으로 본 연구의 제한점과 후속연구를 제시하면 다음과 같다. 첫째, 본 연구에서는 한 교육대학원에서 수행한 교직 적성·인성 검사자료를 사용함으로써 표본이 제한되어 있다. 따라서 모집단을 대표할 수 있도록 교직 적성·인성 검사를 수행하는 다양한 교육대학원의 수학교육 전공 원생들을 대상으로 분석이 수행될 필요가 있다. 둘째, 본 연구에서는 교직 적성·인성 검사지 이외에 발생할 수 있는 오차 요인은 고려되어 있지 않다. 따라서 검사지 이외에 검사의 시행 시기, 시행 횟수, 시행 년도, 시행 시간 등 다양한 오차 요인을 반영한 연구 설계가 필요하다. 셋째, 본 연구에서는 교직 적성·인성 검사를 구성하는 각 영역들이 독립이라는 가정하에 총점을 활용한 일변량 일반화가능도 분석을 수행

하였다. 따라서 하위 영역 간에 상관을 고려한 다변량 일반화가능도 분석을 수행할 필요가 있다. 마지막으로 본 연구에서는 국면의 차원 조정을 통해 적정 수준의 신뢰도에 도달할 수 있는 측정 조건들을 탐색하였지만, 조정된 국면의 내용과 질까지 보장되는 것은 아니다. 즉, 일반화가능도 이론은 전체 검사에 대한 의사결정의 준거만을 제공한다(Keeves, 1998). 따라서 개별 문항에 대한 문항 프로파일분석을 통한 질적인 검토를 수행하는 연구가 필요하다.

참 고 문 헌

강애남, 이규민 (2006). 학생들의 동료평가를 활용한 수행평가 결과의 일반화가능도 분석. 교육평가연구 19(3), 107-121.

Kang, A. N. & Lee, G. M. (2006). A generalizabiliy theory approach to investigating the generalizability of performance assessment using student peer reviews. *Journal of Educational Evaluation 19*(3), 107-121.

교육과학기술부 (2012). 교원자격검정령. 서울: 교육과학기술부.

Ministry of Education, Science, and Technology. (2012). *Official approval provision for teacher qualification.* Seoul: Ministry of Education, Science, and Technology.

교육부 (2013). 교원자격검정령. 서울: 교육부.

Ministry of Education. (2013). *Official approval provisions for teacher qualification.* Seoul: Ministry of Education.

김경령, 서은희 (2014). 예비교사의 교직인성 자기점검 도구 개발 연구. 한국교원교육연구 31(1), 117-139.

Kim, K. R. & Seo, E. H. (2014). The development and validation of the characteristics self-monitoring instrument for pre-service teachers. *The Journal of Korean Teacher Education 31*(1), 117-139.

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

Kim, K. S., Lee, G. M., & Kang, S. H. (2010). Analysis of error sources and estimation of reliability in a Korean Speaking Achievement Test by applying generalizability theory. *Journal of Korean Language Education 21*(4), 51-75.

김민웅, 김태훈 (2017). 공업계열 특성화고 및 마이스터고 학생의 인성 수준 조사 분석. 직업교육연구 36(1), 23-46.

Kim, M. W. & Kim, T. H. (2017). Analysis of personality level of students of industrial-field specialized high schools and Meister high schools. *The Journal of Vocational Education Research 36*(1), 23-46.

김보라, 이규민 (2012). 일반화가능도 이론을 적용한 초등학교 쓰기 수행평가의 총체적 채점과 분석적 채점 방식 비교. 교육학연구 50(4), 49-76.

Kim, B. R. & Lee, G. M. (2012). A comparison of holistic and analytic scoring methods for elementary school writing assessment by applying generalizability theory. *Korean Journal of Educational Research 50*(4), 49-76.

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

Kim, S. S. (1993). The generalizability of student ratings of instructors across sections. *The Journal of Educational Research 31*(1), 23-40.

김성숙 (1998). 준거참조검사의 분할점수에 따른 오차손실 변동과 신뢰성 지수 추정. 교육평가연구 11(1), 153-177.

Kim, S. S. (1998). An estimation of error-loss variation and dependability coefficient associated with cut-off score. *Journal of Educational Evaluation, 11*(1), 153-177.

김성숙, 김양분 (2001). 일반화가능도 이론. 서울: 교육과학사.

Kim, S. S. & Kim, Y. B. (2001). *Generalizability Theory.* Seoul: Kyoyookgwahaksa.

김성찬, 김성연, 한기순 (2012). 관찰, 추전에 의한 수학 영재 선발 시 사용되는 자기소개서와 교사추천서 평가에 대한 일반화가능도 이론의 활용. 수학교육 논문집 26(3), 251-271.

Kim, S. C., Kim, S. Y., & Han, K. S. (2012). An application of generalizability theory to self-introduction letter and teacher's recommendation letter used in identification of mathematical gifted students by observations and nominations. *Communications of Mathematical Education 26*(3), 251-271.

김성호 (2017. 10. 23). 예비교사 인·적성검사, 부적격자 0.6% 그쳐. http://www.shinmoongo.net에서 2017.

10. 23 인출.

Kim, S. H. (2017). *Unqualified candidates were only 0.6% based on aptitude and personality test for pre-service teachers*. Retrieved from http://www.shinmoongo.net.

김성희, 김성연 (2017). 일반화가능도 이론을 적용한 최순자의 유아 사회도덕성 검사의 효율적인 측정 조건 탐색, 미래유아교육학회지 24(1), 325-342.

Kim, S H. & Kim, S. Y. (2017). An investigation of efficient measurement conditions of the sociomoral test for young children by Soon-Ja Choi using generalizability theory, *Journal of Future Early Childhood Education 24*(1), 325-342.

김정환 (2004). 초등학교 기간제 교사의 교육능력 관련 요소의 인과관계 분석, 교육평가연구 17(1), 121-139.

Kim, J. H. (1998). An investigation on casual relationships among educational competence-related factors of period—limit teachers in primary schools, *Journal of Educational Evaluation 17*(1), 121-139.

김정환, 남현우, 염시창, 임진영 (2012). 교직 적성·인성 검사 도구 개발 연구. 서울: 교육과학기술부.

Kim, J. H., Nam, H. W., Yeom, S. C., & Im, J. Y. (2012). *The development of teaching aptitude and personality test*. Seoul: Ministry of Education, Science, and Technology.

김현진 (2013). 교사·학교장 신념과 중학생의 자율성 및 자기효능감, 학업성취도의 관계 분석, 교육학연구 51(2), 117-143.

Kim, H. J. (2013). An analysis of relations among teacher and principal's beliefs and 8th grade students' autonomy, self-efficacy and achievement, *The Journal of Educational Research 51*(2), 117-143.

류춘렬, 이용근 (2010). 일반화가능도 이론을 이용한 집단논리적사고력검사(GALT)의 신뢰도 분석, 한국지구과학회지 31(1), 95-105.

Ryu, C, R. & Lee, Y. G (2010). An analysis of the reliability of Group Assessment of Logical Thinking (GALT) using Generalizability Theory, *The Journal of the Korean Earth Science Society 31*(1), 95-105.

민경석, 박인용, 양길석 (2014). 한국어능력시험Ⅱ의 규준참조적 준거설정 방법 비교, 교육방법연구 26(4), 607-628.

Min, K. S., Park, I. Y., & Yang, K. S. (2014). Evaluation of Norm-referenced Standard Setting Methods for TOPIK Ⅱ, *The Korean Journal of Educational Methodology Studies 26*(4), 607-628.

박정 (2007). 우리나라 중학생의 수학에 대한 정의적 특성 변화와 수학 성취에 미치는 영향력 분석. 수학교육, 46(1), 19-31.

Park, J. (2007). The trend in the Korean middle school students' affective variables toward mathematics and its effect on their mathematics achievements, *The Mathematical Education 46*(1), 19-31.

서경혜, 최진영, 노선숙, 김수진, 이지영, 현성혜 (2013). 예비교사 교직 인성 평가도구 개발 및 타당화, 교육과학연구 26(4), 607-628.

Seo, K. H., Choi, J. Y., No, S. S., Kim, S. J., Lee, J. Y., & Hyun, S. H (2013). The development and validation of teacher disposition assessment instruments, *Journal of Educational Studies 44*(1), 147-176.

신준국, 부덕훈, 서보억 (2015). 수학수업에서 인성 함양을 위한 중학교 교수·학습 자료 개발 연구, 수학교육논문집 29(2), 241-265.

Shin, J. K., Boo, D. H., & Suh, B. E. (2015). A study on the development of teaching and learning materials for character education in middle school, *Communications of Mathematical Education 29*(2), 241-265.

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

Lee, G, M., & Hwang, K. H. (2007). A generalizability theory approach toward investigating the generalizability of scores from holistic and analytic scoring methods in performance assessments of an elementary school science class, *The Korean Journal of Child Education 16*(4), 169-184.

이진아, 한기순 (2016). 일반화가능도 이론을 활용한 TTCT (도형 A형- 활동 2) 독창성 평가 방안 탐색, 창의력교육연구 16(3), 65-77.

Lee, J. A., & Han, K. S. (2016). Optimizing TTCT figure A (Section 2) originality scoring system using the generalizability theory, *The Journal of Creativity Education 16*(3), 65-77.

이태구, 양희원 (2016). 강제결합-스포츠모의중계수업에서 일반화가능도 이론을 적용한 동료평가의 신뢰도

와 오차요인 분석, 체육과학연구 27(2), 345-361.

Lee, T. K. & Yang, H. W. (2016). Analysis of error sources and estimation of reliability in peer review of forced connection method-sportscasting by applying generalizability theory, *Korean Journal of Sport Science 27*(2), 345-361.

이한준, 강민수 (2017). 일반화가능도 이론을 이용한 하지의 등속성 검사의 신뢰도 연구, 운동학 학술지 19(4), 29-35.

Lee, H. J. & Kang, M. S. (2017). Reliability of isokinetic knee strength measurements using generalizability theory, *The Official Journal of the Korean Association of Certified Exercise Professionals 19*(4), 29-35.

이현숙, 송미영 (2015). PISA 2012 수학 성취도를 설명하는 학생의 정의적 특성 및 교사 특성 분석을 위한 다층 구조방정식모형의 적용, 교과교육학연구 19(1), 137-158.

Yi, H. S. & Song, M Y. (2015). A multi-level SEM approach for the analysis of relationships between math-related educational context variables and math literacy of PISA 2012, *Journal of Research in Curriculum Instruction 19*(1), 137-158.

조운주 (2014). 예비유아교사를 위한 교직적성·인성 검사도구의 타당성 및 개선방안, 육아지원연구 9(2), 101-123.

Cho, W. J. (2014). Validation and modification of teaching aptitude test for pre-service early childhood teachers, *Early Childhood Education and Care 9*(2), 101-123.

조주연, 백순근, 임진영, 여태철, 최지은 (2004). 초등 교직적성검사 모형개발 연구, 교육심리연구 18(3), 231-247.

Cho, J. Y., Baek, S. G., Im, J. Y., Yeo, T. C., & Choi, J. E. (2004). The development of the test measuring aptitude for primary school teacher. *Journal of Educational Psychology, 18*(3), 231-247.

조주연, 백순근, 임진영, 여태철, 최지은 (2007). 초등 교직적성검사(TAPST) 타당화 연구, 초등교육연구 20(2), 161-183.

Cho, J. Y., Baek, S. G., Im, J. Y., Yeo, T. C., & Choi, J. E. (2004). A validation study of the Test for Aptitude of Primary School Teacher(TAPST), *The Journal of Elementary Education 20*(2), 161-183.

주지은, 노언경, 이규민 (2007). 공간능력 검사의 성차 및 과제유형 효과와 효율적 측정 구조 탐색, 교육심리연구 21(2), 311-330.

Joo, J. E., No, U. K., & Lee, G. M. (2007). Gender and task type effects on the spatial ability test and the investigation of efficient measurement procedures, *Journal of Educational Psychology 21*(2), 311-330.

한혜숙, 최계현 (2011). 중등 수학 교사들의 정의적 특성에 대한 인식과 수업 실태 분석, 한국학교수학회논문집 14(4), 491-518.

Han, H. S. & Choi, K. H. (2011). Secondary mathematics teachers' recognition of the affective domain and analysis of condition in mathematics teaching, *Journal of the Korean School Mathematics Society 14*(4), 491-518.

한국교육과정평가원 (2016). 역량중심 교육환경에 따른 교사 자격검정 개선 방향. 서울: 한국교육과정평가원.

Korea Institute for Curriculum and Evaluation. (2016). *A study on directions to improve official approval provisions for teacher qualification in a competency-based educational environment*. Seoul: Korea Institute for Curriculum and Evaluation.

홍기칠 (2006). 교직적성·인성 검사도구의 적용연구. 초등교육연구논총 22(1), 113-135.

Hong, K. C. (2006). A Application Study on Test of Aptitude for Primary School Teacher: TAPST. *Journal of Elementary Education, 22*(1), 113-135.

황혜정 (2011). 수학 수업의 교사 지식에 관한 평가 요소 탐색-교수·학습 방법 및 평가를 중심으로, 한국학교수학회논문집 14(3), 241-263.

Hwang, H. J. (2011). The study on the investigation of the mathematics teaching evaluation standards focused on teaching and learning methods and assessment, *Journal of the Korean School Mathematics Society 14*(3), 241-263.

Arce, A. J., & Wang, Z. (2012). Applying Rasch model and generalizability theory to study Modified-Angoff cut scores, *International Journal of Testing 12*(1), 44-60.

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Smith, A. E. (2012). Assessing the dependability of drinking motives via generalizability theory, *Measurement and Evaluation in Counseling and*

Development 45(4), 292-302.

Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk(Ed.), *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: Springer-Verlag.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests, *Educational and Psychological Measurement 55*(2), 157-176.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests, *Journal of Educational Measurement 14*(3), 277-289.

Cizek, G. J. (1996). Standard Setting Guidelines, *Educational Measurement: issues and practice 15*(1), 13-21.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and psychological measurement 20*(1), 37-46.

Englehart, D. S., Batchelder, H. L., Jennings, K. L., Wilkerson, J. R., Lang, W. S., & Quinn, D. (2012). Teacher dispositions: Moving from assessment to improvement, *The International Journal of Educational and Psychological Assessment 9*(2), 26-44.

Fyans, L. J. (1983). *Generalizability theory: Inferences and practical applications*. Jossey-Bass Inc Pub.

Govaerts, M. J., Van der Vleuten, C. P., & Schuwirth, L. W. (2002). Optimising the reproducibility of a performance-based assessment test in midwifery education, *Advances in Health Sciences Education 7*(2), 133-145.

Gugiu, M. R., Gugiu, P. C., & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of grades assigned to undergraduate research papers, *Journal of MultiDisciplinary Evaluation 8*(19), 26-40.

Hanson, B. A. & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models, *Journal of Educational Measurement 27*(4), 345-359.

Haq, R., Anwar, M. N., & Naz, A. (2012). An examination of teaching aptitude of teachers working at primary level: Demographic differences, *International Interdisciplinary Journal of Education 1*(2), 29-33.

Hopkins, K. D. (1997). *Educational and psychological measurement and evaluation* (8th Ed.). Upper Saddle River, NJ: Pearson.

Houchard, M. A. (2005). *Principal Leadership, Teacher Morale, and Student Achievement in Seven Schools in Mitchell County, North Carolina*. Electronic Theses and Dissertations.

Huynh, H. (1976). On consistency of decisions in criterion-referenced testing, *Journal of Educational Measuremen 13*(4), 265-275.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kant, R. (2011). A study of teaching aptitude and responsibility feeling of secondary school teachers in relation to their sex and locale, *Academic Research International 1*(2), 254-259.

Keeves, J. P. (Ed.). (1988). *Educational research, methodology and measurement: An international handbook*. Oxford: Pergamon.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms, *The power of video studies in investigating teaching and learning in the classroom* 137-160.

Kunter, M., Tsai, Y. M., Klusmann, U., Brunner, M.,

Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction, *Learning and Instruction 18*(5), 468-482.

Lee, G. M. (2002). The influence of several factors on reliability for complex reading comprehension tests, *Journal of Educational Measurement 39*(2), 149-164.

Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory, *International Journal of Testing 7*(4), 353-385.

Lin, C. K., & Zhang, J. (2014). Investigating correspondence between language proficiency standards and academic content standards: A generalizability theory study, *Language Testing 31*(4), 413-431.

Livingston, S. A. (1972). Criterion referenced applications of classical test theory, *Journal of Educational Measurement 9*(1), 13-26.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores, *Journal of educational measurement 32*(2), 179-197.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments, *Applied Measurement in Education 15*(3), 249-268.

Narvaez, D., & Nucci, L. P. (2008). *Handbook of moral and character education*. New York, NY: Routledge.

Newton, X. A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis, *Studies in Educational Evaluation 36*(1), 1-13.

Norcini, J. J. (1999). Measurement issues in the use of simulation for testing professionals: Test development, test scoring, standard setting. *Innovative simulations for assessing professional competence*. Chicago, IL: University of Illinois.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan(Ed), *Educational measurement (4th ed)*. Westport, CT: Praeger.

Popham, W. J, & Husek, T. R. (1969). Implications of criterion referenced measurement, *Journal of Educational Measurement 6*(1), 1-9.

Salvia, J., Ysseldyke, J., & Witmer, S. (2012). *Assessment: In special and inclusive education* (11th ed.). Boston, MA: Houghton Mifflin.

Schoonheim-Klein, M., Muijtens, A., Habets, L., Manogue, M., van der Vleuten, C., Hoogstraten, J., & van der Velden, U. (2008). On the reliability of a dental OSCE, using SEM: effect of different days, *European Journal of Dental Education 12*(3), 131-137.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A Primer*. Thousand Oaks, CA: Sage.

Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps(Ed), *Defending standardized testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Solano-Flores, G. & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities, *Educational Measurement: Issues and Practice 25*(1), 13-22.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion referenced test, *Journal of Educational Measurement 13*(4), 265-276.

Sung, Y. T., Chang, K. E., Chang, T. H., & Yu, W. C. (2010). How many heads are better than one? The reliability and validity of teenagers' self-and peer assessments, *Journal of Adolescence 33*(1), 135-145.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion referenced tests: A decision theoretic formulation, *Journal of Educational Measurement 11*(4), 263-267.

Tasleema, J., & Hamid, M. M. (2012). Teaching aptitude of elementary and secondary level teacher educators, *Journal of Education and Practice 3*(2), 67-71.

Taylor, M. A., & Pastor, D. A. (2013). An application of generalizability theory to evaluate the technical quality of an alternate assessment, *Applied Measurement in Education 26*(4), 279-297.

Thompson, A. G. (1992). *Teacher' belief and conceptions: A synthesis of the research*. New York: Macmillan Publishing Company.

Tindal, G., Yovanoff, P., & Geller, J. P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities, *The Journal of Special Education 44*(1), 3-17.

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the Direct Observation Form, *School Psychology Review 38*(3), 382.

Volpe, R. J. & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior, *School Psychology Review 41*(3), 246.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science, *Applied Measurement in Education 13*(3), 277-301.

Yang, Y., Oosterhof, A., & Xia, Y. (2015). Reliability of Scores on the Summative Performance Assessments, *The Journal of Educational Research 108*(6), 465-479.

Yin, P. & Sconing, J. (2008). Estimating standard errors of cut scores for item rating and mapmark procedures: A generalizability theory approach, *Educational and Psychological Measurement 68*(1), 25-41.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use, *Language Testing 24*(2), 251-286.

Investigation of Various Reliability Indices of Pre-service Mathematics Teachers´ Teaching Aptitude and Personality Test based on Setting Cut Scores

Sungyeun Kim
Incheon National University
E-mail : syk@inu.ac.kr

The purpose of this study is first to examine the relative influence of each error source and to investigate the optimal measurement conditions to ensure satisfactory multiple reliability coefficients based on the teaching aptitude and personality test for pre-service teachers. Participants were 33 students enrolled in mathematics education in a graduate school of education located in the Seoul metropolitan area from 2013 to 2017. The main results were as follows. First, the estimated variance due to residual was highest, followed by nesting of items within domains, graduate students, interactions of graduate students with domains, and domains. Second, total 96 items, with 12 domains containing 8 items in each domain, with cut score of 598, and original 210 items, with 14 domains containing 15 items in each domain, with cut scores of 615 or 716 were optimal measurement conditions to reach acceptable reliability levels based on the joint consideration of dependability coefficients, cut score dependability coefficients, adjusted dependability coefficients, and standard errors of measurement. Third, larger deviations between the arithmetic mean and the cut score indicated higher reliability coefficients of the test results. Finally, this study suggests ways for practitioners to consider how to apply generalizability theory for criterion-referenced tests and how to develop future research based on limitations.

* ZDM Classification : C85
* 2000 Mathematics Subject Classification : 97C40
* Key words : teaching aptitude and personality test, dependability coefficients, generalizability theory, criterion-referenced test