韓國數學敎育學會誌 시리즈 E <數學敎育 論文集>
http://dx.doi.org/10.7468/jksmee.2014.28.4.431
제 28집 제 4호, 2014. 11. 431-455

J. Korea Soc. Math. Ed. Ser. E:
Communications of Mathematical Education
Vol. 28, No. 4, Nov. 2014. 431-455

미국 수학교사의 교수 질 평가도구 분석을 통한 우리나라 수학 교원능력개발평가에서의 일반화가능도 이론 활용성 탐색

김 성 연 (서울대학교)

이 연구는 미국 수학교사의 교수 질 평가도구 분석을 통하여 우리나라 수학교사들의 수업관찰 평가의 현장 적용 가능성을 모색하였다. 자료는 2007년부터 미국국립과학재단의 지원을 받아 수행되고 있는 중등수학 교수와 제도적 구성 프로젝트에서 수집한 수학 수업관찰 평가 종단 자료 중 3차년도와 4차년도의 96명의 수학교사 수업관찰 평가점수를 활용하였다. 이 프로젝트는 대규모로 야심차고 공평한 교수 실제를 위한 수학교사의 전문성 개발을 지원하기 위해 필요한 것들을 탐구하고 있다(MIST, 2007). 이 연구에서는 GENOVA 프로그램을 이용하여 단변량 일반화가능도 분석을, 그리고 mGENOVA 프로그램을 이용하여 다변량 일반화가능도 분석을 수행하였다. 구체적으로 교수 질 평가도구를 사용한 수학 수업관찰 평가에서 발생하는 오차요인들의 상대적인 영향력을 살펴보고, 적정 수준의 신뢰도를 확보하기 위한 최적의 측정 조건을 탐색하였다. 이러한 방법론적 틀은 평가의 측정학적 특성을 바탕으로 우리나라 수학교사들의 수업 전문성을 평가하는 교원능력개발평가에서 최적의 측정 조건을 탐구하는데 적용 가능하다. 마지막으로 이 연구의 제한점과 후속연구를 제시하였다.

Ⅰ. 서론

“교육의 질은 교사의 질을 능가하지 못한다”라는 Adam Brooks의 격언처럼, 수학 교과에서도 학생들의 수학 학습에 수학교사의 질은 결정적인 역할을 한다. 실례로, Gläser-Zikuda와 Fuβ(2008)는 학생의 학업성취도를 예측하는 가중 중요한 요인이 교수의 질이며 이는 교사의 역량과 관련되어 있다고 설명하였으며, Hattie(2003)는 교사 효과가 학생의 학업성취도 변량의 30%을 설명, Kyriakides와 Creemers(2008)는 교사 효과가 34%까지 설명, 그리고 Scheerens와 Bosker(1997)는 교사 효과가 학생들의 성취도에 학교와 시스템 수준에서의 변인들보다도 더 높은 설명력을 나타낸다고 밝혔다. 많은 연구들(Cobb & Jackson, 2011; Hiebert & Grouws, 2007; Hill et al., 2008; Hill et al., 2012; Jackson et al., 2013; Learning Mathematics for Teaching Project, 2011; Middle school Mathematics and the Institutional Setting of Teaching Project, 2007; Nye et al., 2004; Rockoff, 2004; Rowan et al., 2002; Teddlie & Reynolds, 2000)은 이러한 수학교사의 질을 측정하기 위해, 우수한 교사들이 전형적으로 잘 가르친다는 관점 하에, 핵심요소를 수학교사의 교수 질로 보고 이를 측정하여 수학교사의 질로 대신하고 있다.

우리나라에서도 교육의 질을 높이기 위한 다양한 노력으로, 수업의 질적 개선에 바탕을 둔 교사의 수업전문성 증진을 목적으로 하는 교원평가 정책을 단계적으로 추진하였다. 2005년-2006년 교원평가의 기본틀 확립과 시

* 접수일(2014년 4월 30일), 심사(수정)일(2014년 7월 28일), 게재 확정일(2014년 8월 12일)
* ZDM 분류 : C73
* MSC2000 분류 : 97C40
* 주제어 : 다변량 일반화가능도 분석, 단변량 일반화가능도 분석, 수학 교수 질 평가
* 이 연구는 2007년부터 2015년까지 ESI-0554535와 DRL-1119112의 재원으로 미국국립과학재단의 지원을 받아 수행된 MIST Project의 자료이며, 자료를 제공해 준 Paul Cobb, Tom Smith, Erin Henrick, 그리고 Anne Garrison Wilhelm을 비롯한 MIST Project의 멤버들에게 감사드립니다.

범학교를 선정하여 운영하고, 교원평가는 교원능력개발평가로 명칭을 변경하였다. 잠시 17대 국회 폐회화 함께 폐기되는 위기를 겪다가, 2008년 선도 시범학교를 전국 30%로 대폭 확대하여 지정하여 운영한 후, 2010년 이후부터 전면적으로 시행되었다(한혜정, 2012). 2011년에는 교원 연수에 관한 규정을 개정하는 등 법령적 근거를 마련하고, 교육행정정보시스템(NEIS, National Education Information System) 연계 온라인평가시스템을 개발하고 보급하였으며, 2012년 교원능력개발평가 실시를 의무화하는 대통령령이 개정되어 현재 5차년도에 접어들고 있다(교육과학기술부, 2013; 김성연, 2014a). 이에 따라 국가 수준에서 한국교육과정평가원의 교수학습개발센터(2014)에서는 우수 수업동영상과 함께 수업관찰 평가에 대한 정보들을 제공하고 있다. 예를 들면, 수학과에서는 관찰 기반의 수업전문성을 평가하는 기준의 의미를 일종의 “좋은 수학 수업”을 정의하는 교수기준으로 규정하여, 이러한 수업전문성을 평가하는 도구로서 수업 관찰지, 교사/학생 면담지, 그리고 전문가 진단지를 어떻게 개발하였으며, 어떻게 측정하고, 또한 어떻게 해석해야 하는지에 대해 제시하고 있다.

이 중 이 연구에서 사용되는 평가도구와 관련이 있는 수업 관찰지에 대해 살펴보면, 수학과 수업전문성 기준의 영역 설정은 이미 개발된 수업평가에 대한 일반 기준(이대현, 최승현, 2006; 임찬빈 외, 2004; Danielson, 1996), 미국수학교사협의회(National Council of Teachers of Mathematics, 1991, 1993)에서 개발한 수학과 수업평가기준, 그리고 수학 수업관찰 및 면담 분석 결과를 참조하여 4개의 대영역, 6개의 중영역, 그리고 평가요소를 나타내는 37개의 문항으로 구성되어 있다. 각각을 살펴보면 다음과 같다.

첫 번째 대영역은 전문적 지식 영역으로 두 개의 중영역이 속해있다. 먼저 수학 교과 지식 및 내용 교수법의 중영역은 교과내용, 다양한 교과 교육, 유용성 관련 내용, 오류 대처 관련 내용, 수업 전략에 대한 지식의 5문항으로 구성되어 있다. 다음으로 학생에 대한 이해의 중영역은 수학 교과 지식 및 내용 교수법의 중영역과 발달과 학습, 개인의 배경 지식과 경험, 학습 방법, 적절한 의사소통능력에 대한 지식의 4문항으로 구성되어 있다.

두 번째 대영역은 계획 영역으로 수업 설계 중영역이 속해있다. 구체적으로 교육과정에 의한 내용 선정, 학습목표에 따른 학습 내용 및 활동 구성, 학생 수준에 따른 수업 내용 구성, 위계성·연계성, 수업 단계 및 학생 수준, 교구 및 자료를 활용한 수업 설계, 학생 평가계획, 그리고 학생 평가결과 활용계획의 8문항으로 구성되어 있다.

세 번째 대영역은 실천 영역으로 두 개의 중영역이 속해있다. 먼저 학습환경 조성 및 학급운영 중영역은 효과적인 수업을 위한 물리적 환경, 수학 학습 문화 조성, 규칙을 통한 학생 관리, 문제 행동 관련 학생 관리하기, 그리고 물리적 환경 유지의 5문항으로 구성되어 있다. 다음으로 수학 수업 실행의 중영역은 학생 선행지식, 학습내용 관련 사전 점검과 동기 유발, 학습 목표와 학습 활동 관련, 학생들에게 유의미한 학습관련 수업 전략, 학습참여도 고취, 학생 자신감과 능력개발, 학생 집단 구성 관련 교사, 효과적인 발문 관련의 수업 운영, 적절한 발문의 피드백 제공, 수업상황에서의 유연한 상황대처, 그리고 평가계획 및 적용의 11문항으로 구성되어 있다.

네 번째 대영역은 전문성 영역으로 전문성 발달의 중영역이 속해 있다. 구체적으로 수업에 대한 자기 반성과 상호 점검, 교과 연구 활동 및 동료 장학, 학부모와의 협력체계 구축, 그리고 전문성 발달을 위한 연구의 4문항으로 구성되어 있다.

이처럼 수학과 수업전문성 기준은 이론적인 대범주를 제안하는 대영역, 실제로 평가기준을 적용하는 관점에서 전문 영역을 고려하여 하위 수준의 영역을 제안하는 중영역이 속해있다. 또한 구체적인 평가기준은 미흡, 기초, 우수, 탁월, 그리고 불가의 5점 척도로 주어지는 문항으로 구성되어 있다(최승현, 임찬빈, 2006).

한편 우리나라에 비해 수학 관련 내용 지식, 수업 포트폴리오, 그리고 부가가치평가 점수 등을 포함한 다양한 평가도구를 활용하여 오랫동안 수학교사의 교수를 측정하기 위해 노력해왔던 미국의 경우, 학교 교장이나 교육행정가들이 교사 효율성을 측정하기 위해 전통적으로 수업관찰 평가를 수행해오고 있다. 그러나 최근 연구자들과 정책 결정자들에 의해 이러한 관찰 기반 평가체계를 개혁해야 한다는 압력이 증가하고 있다. 기존의 교사평

가 제도를 비판하는 위젯효과(The Widget Effect)를 언급한 새 교사 프로젝트(New Teacher Project)의 보고서 발간(Weisberg et al., 2009)과 오바마 정부 출범 후 발표된 정상을 향한 질주(Race to the Top)라는 연방정부의 정책(Jacobs, 2010)에서 예산을 따내기 위한 주정부들의 경쟁이 개혁을 요구하는 압박 요인으로 작용하고 있다. 또 다른 요인으로는 빌게이츠와 멜린다 게이츠가 세운 Bill & Melinda Gates 재단의 효율적인 교수 측정 프로젝트(Measures of Effective Teaching, MET, 2009)를 들 수 있는데, 이 프로젝트에서 발간된 보고서들(Ho & Kane, 2013; Kane & Staiger, 2012)은 교사평가 시 정확한 측정에 대해 강조하며, 다양한 그리고 엄격한 측정도구를 사용할 것을 주장하고 있다. 이러한 요인들은 결과적으로 뉴욕, 메릴랜드, 조지아, 콜로라도, 테네시, 그리고 플로리다를 포함한 많은 주들로 하여금 교실관찰을 기반으로 하는 교사평가 방법을 수정하게 하였다.

실례로 대부분의 주들이 1년에 2번 수학교사들을 관찰(Weisberg et al., 2009)하는 경우와는 다르게, 테네시 주의 경우는 1년에 수학교사들을 4번 관찰(National Center for Teacher Effectiveness, 2011)하던 것을, 초임교원과 경력교원을 고려하여 채점영역에 따라 관찰횟수를 다르게 (Tennessee Department of Education, 2014)하도록 변경하였으며, 반면에 루이지애나는 1년에 오직 1번만 교사들을 관찰(Louisiana Act 54, 2010)하게 하고 있다. 특히 교사평가와 관련하여 연방정부의 정책에서 예산을 따내기 위해서는 부가가치 교사평가 모형을 활용해야한다는 법률이 통과(Eckert & Dabrowski, 2010; 김성연, 2014b)되었는데, 이 모형을 개발한 테네시 주의 경우를 살펴보면, 3개의 채점영역에 19개의 문항을 사용하고 있다.

구체적으로 수업계획, 학생활동, 평가 3문항의 계획영역, 기대, 학생 행동 관리, 교실 환경, 문화존중 4문항의 수업환경영역, 그리고 평가기준과 목표, 동기부여, 수업내용 제시, 수업구조와 시간배분, 활동과 교재, 질문, 학문적 피드백, 그룹 활동, 교수학적 지식, 학생에 대한 지식, 사고, 문제해결 12문항의 교수영역으로 구성되어 있다. 각 문항은 기대이하 1점, 기대수준 3점, 그리고 기대이상 5점의 척도로 구성되어 있으며, 초임교원의 경우 1년에 계획영역과 환경영역의 경우 2번 이상, 그리고 교수영역의 경우 3번 이상 평가를 받게 된다(Tennessee Department of Education, 2014; 김성연, 2014a). 이러한 정책적인 환경 하에서 미국은 수업관찰 평가 체계에 대해 더욱 관심을 기울이고 있으며, 최근 주 교육부, 비영리 연구기관, 그리고 연구자들은 수학과목에서 평가도구 자체뿐만 아니라, 평가점수에 영향을 미치는 채점자, 점수체계, 검사환경 등을 고려하여 신뢰로운 수업관찰 평가 체계를 설계하기 위해 일반화가능도 이론을 적용하고 있다(Colorado Department of Education, 2012; Hill et al., 2012; Kane et al., 2013; Kennedy, 2010; Matsumura et al., 2006, 2008; MET, 2009).

단변량 일반화가능도 분석은 단일오차요인만을 고려하는 고전검사이론의 한계를 보완하며, 연구자가 설정한 측정상황에서 발생하는 복합적인 오차요인을 동시에 분석하여 오차점수에 기여하는 다양한 오차요인을 구분하여 이들을 정량화하고, 이를 통하여 적정수준의 신뢰도에 도달할 수 있는 최적의 측정 절차를 설계할 수 있는 정보를 제공해준다. 다변량 일반화가능도 분석은 관찰 대상자가 여러 개의 점수를 가지고 있으면서, 고정국면 각각과 관련된 임의효과 분산 성분을 가지도록 설계하여 분석하는 방법으로, 단변량 일반화가능도 분석은 다변량 일반화가능도 분석의 특수한 형태이다.

단변량 일반화가능도 분석에 비해 다변량 일반화가능도 분석을 사용하는 이유는 첫째, 고정국면과 임의국면이 혼합된 모형이면서 불균형 설계인 경우에 단변량 일반화가능도 분석을 실시하면 매우 복잡하지만, 다변량 일반화가능도 분석은 모든 채점영역의 평가 결과를 동시에 분석할 수 있어 상대적으로 쉽게 적용할 수 있기 때문이다. 둘째, 단변량 일반화가능도 분석에서 제공해주지 못하는 각 고정효과 국면의 채점영역과 관련되어 있는 임의효과 국면의 분산, 공분산 성분 추정치를 제공해주며, 이를 바탕으로 고정효과 국면의 수준별 전집점수에 가중치를 준 합성점수에 대해 오차 분산 및 신뢰도를 제공해주기 때문이다(김성숙, 2001; 김성연, 한기순, 2013, 2014; 이현숙, 2012; Brennan, 2001; Cronbach et al., 1972; Webb et al., 1983).

따라서 이 연구의 목적은 수학교사의 수업을 관찰하는 평가도구에 대해 오랜 연구가 이루어진 미국의 실제 현장에서 사용된 수업관찰 평가 자료를 활용하여 측정 조건에 따른 신뢰도의 변화를 바탕으로 해당 평가 자료

에 대한 효율적인 측정 조건을 탐색하는 방법을 예시함으로써 우리나라 수학교육 현장에서의 적용 가능성을 탐구하는 데 있다. 예시 자료인 교수 질 평가(Instructional Quality Assessment, IQA) 도구(Jackson, et al., 2013; MIST, 2007)를 활용하여 측정한 미국의 수학 수업관찰 평가 자료에 일반화가능도 분석을 활용하여 밝히고자하는 이 연구의 구체적인 연구문제는 다음과 같다.

첫째, IQA 도구를 활용한 수학 수업관찰 평가에서 서로 다른 국면들을 고려함에 따라 교사의 평가점수에 영향을 주는 요인들의 상대적 영향력은 어떻게 변화하는가?

둘째, IQA 도구를 활용한 수학 수업관찰 평가에서 서로 다른 국면들을 고려하는 연구 설계에 따라 적정수준의 신뢰도를 얻기 위한 최적의 측정조건은 무엇인가?

Ⅱ. 이론적 배경

1. 교수 질 평가(Instructional Quality Assessment) 도구

수학 수업관찰 평가에 사용되는 도구는 연구자들에 따라 평가도구 이름, 채점영역, 그리고 채점영역에 속하는 문항들의 이름은 달라지지만, 공통적으로 일부 문항에 대해서는 일관된 채점기준이 적용되고 있다. 또한 해를 거듭함에 따라 기존 평가도구의 신뢰도 분석 결과에 따라, 또는 연구자들의 필요에 의해 수정·보완된 평가도구들이 학회와 학회지를 통해 계속 발표되고 있다. 따라서 이 연구에서는 IQA 도구라는 이름을 갖는 평가도구에 한해 각 연구자들마다 최신 연구에서 소개한 서로 다른 3종류의 IQA 도구에 대해 연도별로 Matsumura 외(2006), Boston(2012), 그리고 Jackson 외(2013)의 루브릭에 대해 살펴보았다. 각 IQA 도구에서 문항의 채점기준이 같은 경우는 비록 채점영역이나 문항의 이름이 다른 경우에도 이 연구에서는 같게 취급하여, Matsumura 외(2006)를 기준으로, 추가되는 채점 기준만을 따로 제시하였다.

Matsumura 외(2006)는 책무성 대화영역, 학문적 엄격성영역, 그리고 명확한 기대영역 총 3개 영역으로 구성되어 있다. Matsumura 외(2006)가 개발한 IQA 도구의 루브릭에 대한 예시는 <표 1>에 제시되어 있다.

<표 1> Matsumura 외(2006)가 개발한 IQA 도구의 루브릭

채점영역	문항	배점	세부 기준
책무성 대화	참여 B, J	4	토론에 참여한 학생들의 비율이 75% 이상
		3	토론에 참여한 학생들의 비율이 50-75%
		2	토론에 참여한 학생들의 비율이 25-49%
		1	토론에 참여한 학생들의 비율이 25% 미만
		0	토론에 참여한 학생들이 없음
	교사 연계 B, J	4	교사는 지속적으로 학생들의 응답들을 서로에게 연결시키고, 학생들의 아이디어를 반향하거나 재현함으로써 토론 중에 공유된 아이디어들이 서로 어떻게 관련되어 있는지 보여줌
		3	교사는 수업 중 적어도 두 번 학생들의 응답들을 서로에게 연결시키고, 학생들의 아이디어를 반향하거나 재현함으로써 토론 중에 공유된 아이디어들이 서로 어떻게 관련되어 있는지 보여줌
		2	교사는 토론 중 한 번 이상 학생들의 응답들을 서로에게 연결시키지만, 아이디어

			들이 서로 어떻게 관련되어 있는지는 드러나지 않음
		1	교사는 학생들의 응답들을 연결시키거나 반향하는 노력을 전혀 하지 않음
		0	학급 토론은 수학과 관련이 없음
책무성 대화	학생 연계 B, J	4	학생들은 지속적으로 그들의 대답들을 서로에게 연결하고, 토론 중에 공유된 아이디어들이 서로 어떻게 관련되어 있는지를 보여줌
		3	학생들은 수업 중 적어도 두 번 그들의 대답들을 서로에게 연결하고, 토론 중에 공유된 아이디어들이 서로 어떻게 관련되어 있는지를 보여줌
		2	학생들은 토론 중 한 번 이상 그들의 대답들을 서로에게 연결하지만, 아이디어들이 서로 어떻게 관련되어 있는지는 드러나지 않음
		1	학생들은 다른 학생들의 대답을 연결시키거나 반향하는 노력을 전혀 하지 않음
		0	학급 토론은 수학과 관련이 없음
	교사 발문 B, J	4	교사는 지속적으로 학생들에게 그들의 응답에 대한 증거를 제공, 정확성을 요구, 또는 그들의 추론을 설명하도록 유도할 수 있는 질문들을 함
		3	교사는 수업 중 적어도 두 번 학생들에게 그들의 응답에 대한 증거를 제공, 정확성을 요구, 또는 그들의 추론을 설명하도록 유도할 수 있는 질문들을 함
		2	교사는 한 번 이상 학생들에게 그들의 응답에 대한 증거를 제공, 정확성을 요구, 또는 그들의 추론을 설명하도록 피상적, 사소한, 그리고 단순한 노력만을 함
		1	교사는 학생들에게 그들의 응답에 대한 증거를 제공하도록, 그리고 그들의 사고를 설명하도록 요구하는 어떤 노력도 하지 않음
		0	학급 토론은 수학과 관련이 없음
	학생 응답 B, J	4	학생들은 지속적으로 교재나 이전의 교실 경험을 바탕으로 그들의 주장에 대해 정확하고 적절한 증거를 제공, 또는 적절한 추론을 활용하여 자신들의 생각을 설명함
		3	학생들은 수업 중 적어도 두 번 교재나 이전의 교실 경험을 바탕으로 그들의 주장에 대해 정확하고 적절한 증거를 제공, 또는 적절한 추론을 활용하여 자신들의 생각을 설명함
		2	전반적으로 학생들이 그들의 주장을 뒷받침하기 위해 제공하는 증거들은 부정확, 불완전, 애매하며, 또는 증거를 제공하기 위해 한 번 이상의 피상적이거나 사소한 노력만이 있음
		1	학생들은 그들의 주장을 지지하지 못하며, 또는 그 이유를 설명하지 못함
		0	학급 토론은 수학과 관련이 없음
학문적 엄격성	과제 가능성 B, J	4	과제는 학생들이 수학적 개념, 절차, 그리고/또는 관계의 본질을 잠재적으로 이해하고 탐색하도록 함
		3	과제는 학생들이 복잡한 사고 또는 수학적 개념, 절차, 그리고/또는 관계의 의미를 잠재적으로 형성하도록 하지만, 4점에는 미치지 못함
		2	과제의 핵심은 수학적 이해를 발전시키기보다는 정답을 맞추는데 있음
		1	과제는 학생들이 사실, 규칙, 공식 또는 정의를 암기하거나 재현하는데 한정됨
		0	과제는 어떤 수학적 활동도 요구하지 않음

	과제 수행 B, J	4	학생들은 수학적 개념, 절차, 그리고/또는 관계의 본질을 이해하고 탐색하는데 참여함
		3	학생들은 복잡한 사고 또는 수학적 개념, 절차, 그리고/또는 관계의 의미를 형성하는데 참여하지만, 4점에는 미치지 못함
		2	과제 수행의 핵심은 수학적 이해를 발전시키기보다는 정답을 맞추는데 있음
		1	학생들은 사실, 규칙, 공식 또는 정의를 암기하거나 재현하는데 참여함
		0	과제는 어떤 수학적 활동에도 참여하지 않음
	과제 이후의 학생토론 B, J	4	학생들은 왜 그들의 전략, 아이디어 또는 절차가 타당한지를 완벽하게 설명하고, 글로써 설명/묘사함
		3	학생들은 글로써 설명/묘사할 수 있지만, 설명은 불완전함
		2	학생들은 과제 해결을 글로써 설명/묘사할 수 있지만, 왜 그들의 전략, 아이디어 또는 절차가 그 문제에 적합한지는 설명하지 못함
		1	학생들은 간단하거나 괄호 채우기 문제에 대한 정답만 제공함
		0	학생들의 응답은 수학과 관련이 없음
명확한 기대	기대의 명확성과 세부사항 B	4	교사는 학생들에게 과제를 잘 수행하기 위해 그들이 해야 할 것, 또는 그들의 그룹 활동에 포함되어야 할 것들에 대해 명확한 설명을 제공, 수준 높은 학생들의 그룹 활동을 예시, 그리고/또는 수준에 따른 그룹 활동의 차이를 묘사함
		3	교사는 학생들에게 과제를 잘 수행하기 위해 그들이 해야 할 것, 또는 그들의 그룹 활동에 포함되어야 할 것들에 대해 명확한 설명을 제공함
		2	교사는 학생들의 그룹 활동의 질에 대해 기대하는 정도에 대해 피상적 또는 일반적인 설명을 제공함
		1	교사는 그룹 활동에 대한 방향을 제시하지만, 과제를 잘 수행하기 위해 학생들이 해야 할 것, 또는 그들의 그룹 활동에 포함되어야 할 것들에 대해 설명하지 않음
		0	교사는 그룹 활동의 질에 대한 기대를 학생들과 공유하지 않음

주. B는 Boston(2012)이, J는 Jackson 외(2013)가 개발한 IQA 도구에 포함되는 문항들임

Boston(2012)은 과제 가능성의 1개 문항으로 구성된 교육적 과제 영역, 수행의 1개 문항으로 구성된 과제 수행 영역, 토론의 엄격성, 참여, 교사 연계, 학생 연계, 교사 압력, 학생 응답의 6개 문항으로 구성된 수학적 사고와 추론 설명 영역, 그리고 교사 기대의 엄격성, 명확성과 세부사항, 학생 접근성의 3문항으로 구성된 교사의 기대 영역으로 총 3개 영역의 11 문항으로 구성되어 있다. Boston(2012)이 개발한 IQA 도구의 루브릭 중 Matsumura 외(2006)가 개발한 문항들과 중복되지 않는 문항들의 채점기준은 <표 2>에 제시되어 있다.

<표 2> Boston(2012)이 개발한 IQA 도구의 루브릭

채점영역	문항	배점	세부 기준
교사의 기대	교사 기대의 엄격성	4	교사 기대의 대부분은 학생들이 복잡하고, 비알고리즘적인 사고를 활용하며, 수학적 개념, 절차 그리고/또는 관계의 본질을 탐색하고 이해하기를 바람
		3	교사 기대의 적어도 일부는 학생들이 복잡한 사고 또는 중요한 수학적 개념을 이해하기를 바라지만, 4점에는 미치지 못함
		2	교사의 기대는 학생들이 학습과 관련된 기술에 초점을 맞추고 있지만, 이것들은 복잡한 사고 기술은 아님
		1	교사의 기대는 실제적인 수학 개념에 초점을 맞추지 않음
		0	교사는 학생들에게 기대를 표현하지 않음
	학생 접근성	4	기대되는 그룹 활동의 질에 대한 채점 기준과 점수체계는 모든 학생들이 쉽게 열람할 수 있으며, 전체 학급을 대상으로 이러한 채점 기준을 발표함
		3	기대되는 그룹 활동의 질에 대한 채점 기준은 모든 학생들에게 설명하지만, 전체 학급을 대상으로 이러한 채점 기준을 발표하지 않음
		2	기대되는 그룹 활동의 질에 대한 채점 기준은 일부 학생들에게 설명하며, 전체 학급을 대상으로 이러한 채점 기준을 발표하지 않음
		1	그룹 활동의 질에 대한 기대를 학생들과 공유하지 않음
		0	그룹 활동은 수학과 관련이 없음

Jackson 외(2013)는 과제 가능성, 과제 수행의 2개의 문항으로 구성된 인지적 요구 영역, 토론의 학문적 엄격성, 참여, 교사 연계, 학생 연계, 교사 발문, 학생 응답의 6개 문항으로 구성된 토론 영역, 그리고 문맥 특성, 수학적 관계, 인지적 요구 유지, 셋업 후 과제 특성, 셋업 참여의 5개 문항으로 구성된 셋업 영역으로 총 3개 영역의 13개 문항으로 구성되어 있다. Jackson 외(2013)가 개발한 IQA 도구의 루브릭 중 Matsumura 외(2006)와 Boston(2012)이 개발한 문항과 중복되지 않는 문항들의 채점기준은 <표 3>에 제시되어 있다.

<표 3> Jackson 외(2013)가 개발한 IQA 도구의 루브릭

채점영역	문항	배점	세부 기준
셋업	맥락적 특성	4	교사 그리고/또는 학생들은 지속적으로 문제 해결 시나리오의 맥락적 특성의 공유된 이해 확립을 위해 지지 그리고/또는 구축하는 아이디어들을 연결함
		3	교사 그리고/또는 학생들은 비지속적으로 문제 해결 시나리오의 맥락적 특성의 공유된 이해 확립을 위해 지지 그리고/또는 구축하는 아이디어들을 연결함
		2	교사는 학생들이 문제 해결 시나리오에 대해 알고 있는 것들을 이끌어내지만, 교사 그리고/또는 학생들은 문제 해결 시나리오의 맥락적 특성의 공유된 이해 확립을 도와주는 아이디어들을 연결시키지 못함
		1	교사는 적어도 과제를 완수하는데 중심이 되는 문제 해결 시나리오를 간략하게 언급함
		0	문제 해결 시나리오의 맥락적 특성에 대해 토론하려는 시도가 없음
	수학적 관계	4	교사는 학생들이 발전시킨 아이디어를 이끌어내고, 과제에 나타난 주요 수학적 아이디어, 관계 그리고/또는 수치들에 대해 공유된 이해를 확립하도록, 적어도 한 번 이상의 강한 책무성 대화가 나타나야함
		3	교사는 학생들이 발전시킨 아이디어에 대한 정보를 이끌어내고, 과제에 나타난 주요 수학적 아이디어, 관계 그리고/또는 수치들에 대해 공유된 이해를 확립하도록 지속적으로 책무성 대화를 사용함
		2	교사는 학생들이 발전시킨 아이디어에 대한 정보를 이끌어내지만, 과제에 나타난 주요 수학적 아이디어, 관계 그리고/또는 수치들에 대해 공유된 이해를 확립하도록 기껏해야 비지속적으로 책무성 대화를 사용함
		1	교사는 적어도 어떻게 주요 수학적 아이디어, 관계 그리고/또는 수치들이 과제에

			나타나는지를 제시하고자 시도를 함
		0	주요 수학적 아이디어, 관계 그리고/또는 수치들에 대해 토론하려는 시도가 없음
	셋업 유지	1	과제 엄격성이 증가하거나 유지됨
		0	과제 엄격성이 감소함
	셋업 후 과제 특성	4	과제는 학생들이 수학적 개념, 절차, 그리고/또는 관계의 본질을 잠재적으로 이해하고 탐색하도록 했음
		3	과제는 학생들이 복잡한 사고 또는 수학적 개념, 절차, 그리고/또는 관계의 의미를 잠재적으로 형성하도록 하지만, 4점에는 미치지 못했음
		2	과제의 핵심은 수학적 이해를 발전시키기보다는 정답을 맞추는데 있었음
		1	과제는 학생들이 사실, 규칙, 공식 또는 정의를 암기하거나 재현하는데 한정되었음
		0	과제는 어떤 수학적 활동도 요구하지 않았음
	셋업 참여	4	셋업에 참여한 학생들의 비율이 75% 이상
		3	셋업에 참여한 학생들의 비율이 50-75%
		2	셋업에 참여한 학생들의 비율이 25-49%
		1	셋업에 참여한 학생들의 비율이 25% 미만
		0	셋업에 참여한 학생들이 없음

2. 일반화가능도 이론을 적용한 교사 평가

교사를 측정의 대상으로 삼아 일반화가능도 이론을 적용한 연구들은, 교사나 교수들을 채점자로 하고, 학생을 측정의 대상으로 삼는 과목별 성취도 평가, 수행평가, 영재선발 도구, 임상수행능력평가 그리고 체육학연구 등(김경선 외, 2010; 김도연, 허종관, 2002; 김보라, 이규민, 2012; 김성숙, 2006; 김성연, 한기순, 2013, 2014; 김성찬 외, 2012; 김현철, 2003; 이규민, 황경현, 2007; 이향, 2012; 임형, 김성숙, 2005; 조재윤, 2009; 최숙기 2012)에 비해 상대적으로 훨씬 적은 편이다. 실례로 국내의 경우 교사를 측정의 대상으로 삼아 단변량 일반화가능도 이론을 적용한 연구는 김성숙(1989a 1989b, 1992)의 연구가, 그리고 다변량 일반화가능도 이론을 적용한 연구는 김성연(2014)의 연구가 유일하다. 김성숙(1989a, 1989b, 1992)의 연구는 모두 미국 버지니아 주에서 실시한 초임교사 지원프로그램(The Beginning Teacher Assistance Program, BTAP)을 위한 채점자 훈련과정에 참가한 교사들이 고등학교 영어 시간과 초등학교 사회시간이 녹화된 다른 교사의 비디오를 보고, 루브릭에 나타난 문항에 대해 교사의 행동 여부를 기록한 것을 자료로 사용하였다. 이 연구들(김성숙, 1989a, 1989b)은 관찰을 통하여 교사 행동을 측정하는데 있어 교사에 내재된 관찰횟수를 나타내는 경우(occasion)에 의한 분산의 크기가 상대적으로 크게 나타남으로써 교사 행동을 여러 번 관찰하여야 함을 제안하였다. 또한 후속연구(김성숙, 1992)에서는 채점자 경험이나 배경이 교사 행동을 관찰하는데 있어 측정의 오차원에 기여하지는 않으며, 측정기록의 신뢰도에 미치는 영향력에 있어 채점자에 의한 변동보다는 관찰횟수를 나타내는 경우(occasion) 효과가 더 크게 나타났다며 사전 연구와 일치하는 결과를 제시하였다.

한편 김성연(2014)은 미국 남부의 한 주(state)에서 실시한 수업관찰 평가 결과 중 주(state)에서 실시하는 채점자훈련과정을 모두 마치고, 채점인증시험에 합격한 채점자들에 의해 평가점수가 부여된 총 3847명의 초임교사 자료를 활용하여 관찰횟수가 신뢰도에 미치는 상대적인 영향력과 최적의 측정 조건을 탐색하였다. 그 결과 초임교원의 유형에 상관없이 측정의 대상인 교사 분산이 크게 나타났으며, 관찰횟수 분산은 작게 나타났다고 밝혔다. 또한 문항 수를 연구에 사용된 평가도구와 같게 한 상태에서 관찰횟수를 조정해가며 적정 수준의 신뢰도를 얻기 위한 최적의 측정 조건을 탐색한 결과, 현재 실시하고 있는 3번의 교수영역, 2번의 수업환경영역, 그리고 계획영역은 각각 교수영역은 2번, 계획영역은 1번으로 감소, 그리고 수업환경영역은 동일하게 2번 관찰할 것을 제시하였다.

반면에 국외에서는 수학교사들의 수업관찰 평가와 관련하여 일반화가능도 이론을 적용한 연구들이 최근 활발히 진행되고 있다. Matsumura 외(2008)는 교사들의 읽기와 수학 수업관찰, 그리고 읽기와 수학 과제물을 바탕으로 교수 질을 평가하는데 단변량 일반화가능도 분석을 적용하였다. 이 연구에서는 수학수업 관찰에만 초점을 맞추어 살펴보면, 총 11명의 교사(t)들에게 2번의 수업관찰(o)을 통한 평가점수가 부여되는 $t \times o$ 설계가 적용되었다. G-연구 결과, 측정의 대상인 교사 분산이 가장 크게 나타났으며, 관찰횟수 분산은 거의 존재하지 않는 것으로 나타났다. G-연구결과를 바탕으로 수행한 D-연구 결과, 관찰횟수가 2번인 경우 절대평가에 사용되는 신뢰도인 의존도 계수가 0.86, 관찰횟수가 3번, 4번 그리고 5번으로 증가함에 따라 각각 0.90, 0.92 그리고 0.94로 각각 증가하는 것으로 나타났다. 그러나 이 연구에서는 관찰횟수 분산이 0으로 나타남에 따라 실제 상대평가에 사용되는 일반화가능도계수도 같은 결과를 산출함을 알 수 있다. 이러한 연구결과는 IQA 도구를 활용하여 신뢰로운 평가점수를 얻기 위해서는 절대평가나 상대평가 모두 2번의 관찰만으로도 충분하다는 결론을 제시하였다.

Hill, Charalambous, 그리고 Kraft(2012)는 수학교사를 평가함에 있어 관찰평가가 성공적으로 수행되기 위해서는 관찰평가 체계가 발전되어야 하는데, 이러한 체계에는 관찰도구 자체에 대한 내용뿐만 아니라 적절한 채점자 선정과 채점자 훈련기간을 통해 신뢰롭고 효율적인 관찰평가 점수를 산출할 수 있는 설계가 필요하며, 바로 이 부분에 일반화가능도 이론을 적용해야 한다고 실제 예를 들어 설명하였다. 구체적으로 수학을 가르치는데 필요한 수학적 지식의 수준이 다른 8명의 수학교사를 대상으로, 녹화된 수업을 7.5분씩 구분한 총 24개의 수업내용을 9명의 채점자가 수학 교수 질(Mathematical Quality of Instruction, MQI) 도구(Hill et al., 2008)로 채점한 자료를 사용하였다. MQI는 수학적 풍부함(richness, R) 영역, 교사 실수와 부정확성(errors and imprecision, EI) 영역, 그리고 수학적 의미 결정과 추론에 학생들의 참여(the student participation in meaning making and reasoning, SPMMR) 영역의 3영역으로 구성되어 있다. R영역에는 5개의 문항이, EI와 SPMMR영역에는 3개의 문항이 포함되어 있으며, 이 연구에서는 영역별로 단변량 일반화가능도 분석을 적용하였다.

구체적으로 24개의 수업내용(l)은 관찰된 시점에 따라 교사들마다 다를 것이므로, 3개의 채점영역에 8개의 수업내용이 내재되어 있으며, MQI의 각 영역별로 9명의 채점자(r)가 8명의 수학교사(t)를 모두 관찰한 평가점수가 주어지는 $(l:t) \times r$설계가 적용되었다. 먼저 G-연구 결과, 5문항의 평균을 이용한 R영역에서는 교사에 의한 분산이 가장 크게 나타났으며, 다음으로 잔차, 교사에 내재된 수업내용, 교사와 채점자의 상호작용 그리고 채점자의 순으로 나타났다. 3문항의 평균을 이용한 EI 영역에서는 잔차에 의한 분산이 가장 크게 나타났으며, 다음으로 교사, 채점자, 교사에 내재된 수업내용, 그리고 교사와 채점자의 상호작용의 순으로 나타났다. 또한 3문항의 평균을 이용한 SPMMR영역에서는 교사에 의한 분산이 가장 크게 나타났으며, 다음으로 잔차, 채점자, 교사에 내재된 수업내용, 그리고 교사와 채점자의 상호작용의 순으로 나타났다.

G-연구결과를 바탕으로 한 D-연구결과에서는 전체 수업과 수업의 첫 30분만을 수업내용으로 간주하여 각 채점영역 별로 수업내용 수와 채점자수를 조정하며 구한 일반화가능도계수를 제시하였다. 전체수업을 수업내용으로 고려하는 경우, 적정수준인 0.80의 일반화가능도계수를 얻을 수 있는 최적의 측정 조건은 R 영역에서는 3번의 수업내용과 3명의 채점자 또는 4번의 수업내용과 2명의 채점자, EI 영역에서는 3번의 수업내용과 4명의 채점자 또는 4번의 수업내용과 3명의 채점자, 그리고 SPMMR 영역에서는 2번의 수업내용과 4명의 채점자, 또는 3번의 수업내용과 2명의 채점자가 필요한 것으로 나타났다. 반면에 첫 30분만을 수업내용으로 고려하는 경우, 적정수준인 0.80의 일반화가능도계수를 얻을 수 있는 최적의 측정 조건은 R 영역에서는 3번의 수업내용과 2명의 채점자, EI 영역에서는 4번의 수업내용과 4명의 채점자, 그리고 SPMMR 영역에서는 4번의 수업내용과 4명의 채점자를 고려하여도 0.80에는 미치지 못하는 것으로 나타났다.

또한 어떻게 효율적인 교수방법이 확인되고 개발될 수 있는지를 연구하는 MET 프로젝트에서 나온 보고서에 따르면, 수학수업 관찰의 경우, 교수를 위한 구조(Framework for Teaching), 교실평가점수체계(Classroom Assessment Scoring System), MQI, 그리고 교사관찰 프로토콜(UTeach Teacher Observation Protocol) 도구들

을 사용하고 있다. 구체적으로 Kane과 Staiger(2012)은 수업관찰 평가에서 동일한 교사에 대해 채점자가 누구인지, 그리고 수업내용에 따라 관찰점수에 차이가 나므로 신뢰도를 높이기 위해서는 각 교사들에 대해 여러 번의 수업내용을 여러 명의 채점자가 평가할 것을 제시하였다.

또한 이 연구의 후속연구로 Ho와 Kane(2013)은 자발적으로 지원한 67명의 교사에 대해 수업하는 장면을 녹화하고, 이 자료를 활용하여 일반화가능도 이론을 적용한 6가지 연구 결과를 제시하였다. 그 중 김성연(2014)에서 보고한 두 가지의 연구결과를 제시하면 다음과 같다. 먼저 녹화된 비디오를 교장집단(해당 교사가 다니는 학교의 교장과 다른 학교 출신의 교장)과 동료교사집단(같은 학년을 담당하는 교사와 다른 학년을 담당하는 교사)으로 나누어 평가 받아야 할 비디오를 교사 본인이 직접 선정하는 경우와 그렇지 않은 경우를 비교한 결과, 교사 본인이 직접 선정한 비디오의 평가점수가 높게 나타났지만, 전체 교사들의 상대적인 순위가 바뀌지는 않는다고 했다. 오히려 교사 본인들이 선정한 비디오의 평가점수의 편차가 크게 나타남으로써 더 변별력이 있다고 해석할 수 있으므로, 교사들에게 관찰평가의 부담감을 줄여주는 방안으로 평가받을 시점과 수업 내용을 교사 본인들이 정할 수 있도록 교사들에게 선택권을 줄 것을 제안하였다. 또한 적정 수준의 신뢰도를 얻기 위한 최적의 측정 조건으로 4개의 시나리오를 분석한 결과 수업 전체를 녹화한 60분 비디오를 1명의 채점자가 평가하는 것보다는 60분 비디오를 15분으로 나누어서 4명의 채점자가 평가할 것을 관찰평가에서 효율적인 측정 조건으로 제안하였다.

이상의 선행연구에서 살펴본 바와 같이, 미국에서는 최근 수학수업의 질을 향상시키기 위한 노력의 일환으로 수학교사의 수업을 녹화한 후, 이를 여러 가지 수업관찰 평가도구를 활용하여 수학교사의 교수 질을 측정하고 있다. 따라서 이러한 측정에 앞서 우선시 되어야 할 부분은 평가도구에 대한 객관성의 확보가 중요하며, 측정학적 이론을 바탕으로 평가도구의 신뢰도를 확보할 수 있는 구체적인 방안을 모색하기 위해, 일반화가능도 이론이 적용되고 있음을 알 수 있다. 그러나 일반화가능도 이론은 주어진 연구대상과 고려하는 오차 국면에 따른 연구 설계 등에 따라 다른 결과들을 나타낼 수 있다. 따라서 이 연구에서는 하나의 자료를 바탕으로 다양한 연구 설계를 고려함으로써 오차국면의 효과와 최적의 측정조건을 탐색하는 방법을 예시하였다.

Ⅲ. 연구 방법

1. 자료수집

이 연구는 2007년부터 미국국립과학재단(The National Science Foundation, NSF)의 지원으로 수행되고 있는 중등수학 교수와 제도적 구성(Middle-school Mathematics and the Institutional Setting of Teaching, MIST, 2007) 프로젝트 종단자료 중 일부를 활용하였다. MIST 프로젝트는 대규모로 야심차고 공평한 교수 실제를 위한 수학 교사의 전문성 개발을 지원하기 위해 필요한 것들을 탐구하고 있다. MIST 1단계에는 2007년부터 2011년까지 4개의 큰 도시 지역의 360,000명 중학생(12세-14세), 지역 별로 6-10개 학교, 각 지역 별 30명의 중등수학교사, 그리고 15-20명의 수학코치, 교장, 그리고 지역 리더들이 참가하였다. 매 해 10월, 각 지역의 지도자들이 중등수학 수업을 향상시키기 위해 지원하는 전략들을 조사하는 인터뷰가 시행되었고, 1월부터 3월까지는 MIST 연구진들이 실제 학교와 교실 현장에서 이러한 전략들이 어떻게 수행되고 있는지를 문서화하였다. 수집한 자료에는 교사와 수업 지도자가 근무하는 학교와 지역 내에서 수행된 인터뷰 오디오 녹음, 교사, 코치 그리고 교장들의 온라인 설문조사, 교사들의 수학수업 비디오 녹화, 교수를 위한 수학지식(Mathematics Knowledge for Teaching, MKT) 도구(Hill et al., 2008; Learning Mathematics for Teaching Project, 2011)로 평가한 교사와 코치의 점수, 선별한 지역 전문 개발가의 인터뷰 비디오 녹화, 교사 협동 계획 회의의 오디오 녹음, 교사 네트워크의 온라인 평가, 그리고 학생들의 수학 성취도 자료가 포함되어있다(Cobb & Jackson, 2011).

<표 4> IQA 도구의 채점영역과 문항내용

채점영역	문항	수업 관찰시 살펴보아야 할 중요한 측면	배점
인지적 요구	과제 가능성	교육과정 자료에 제시되어 있는 과제의 인지적 요구 수준	0-4
	과제 수행	학생들이 과제를 해결하기 시작하면서부터 수업 마칠 때까지 실제 수행된 인지적 요구 수준	
토론	토론의 학문적 엄격성	전체 학급 토론에서의 학문적 엄격성	
	참여	전체 학급 토론에 참여한 학생들의 비율	
	교사 연계	전체 학급 토론 내에 기여하는 교사 연계 부분	
	학생 연계	전체 학급 토론 내에 기여하는 학생 연계 부분	
	교사 발문	전체 학급 토론 내에서 개념적 설명을 위한 교사 발문	
	학생 응답	전체 학급 토론 내에서 학생이 제공하는 개념적 설명	
셋업	문맥 특성	과제 진술에서 문제 해결 시나리오의 문맥 특성에 대해 공유된 이해 구축	
	수학적 관계	과제 진술에서 수학적 관계와 아이디어에 대해 공유된 이해 구축	
	인지적 요구 유지	셋업 단계 내내 구체적인 과제의 인지적 요구 수준 유지	0-1
	셋업 후 과제 특성	셋업 마지막단계에서 교수자료를 근거로 과제의 인지적 요구 수준 유지	0-4
	셋업 참여	셋업 토론에 참가한 학생들의 비율	

이 연구에서는 MIST 프로젝트의 3차년도와 4차년도 자료 중 Matsumura 외(2006)가 개발한 11개 문항 중 8개 문항과 MIST 프로젝트에서 자체 개발한 5개 문항을 합친, 3개 영역 13개 문항으로 구성된 IQA 도구를 활용하여, 비디오로 녹화된 수학교사의 수업관찰 평가 자료를 사용하였다. IQA 도구의 채점영역과 문항내용은 <표 4>와 같다. 구체적으로 총 150명의 교사 중, 96명만이 인지적 요구영역, 토론영역, 그리고 셋업영역의 3개 영역 모두에서 이틀 동안 관찰되었으며, 41명은 토론영역에서 이틀 중 하루만, 그리고 13명은 토론영역의 문항들이 이틀 동안 모두 관찰되지 않았다. 따라서 이 연구에서는 4가지 연구 설계 결과의 비교를 용이하게 하기 위해, Matsumura 외(2008)의 연구와 마찬가지로 수업관찰 이틀 동안 토론영역에 결측치가 모두 없는 96명의 수학교사만으로 연구대상을 한정하였다.

2. 자료분석

이 연구에서는 4가지 연구 설계에 포함된 국면에 따라 채점영역인 인지적 요구영역, 토론영역 그리고 셋업영역은 고정된 국면으로 정의하고, 채점영역 내의 세부영역을 나타내는 문항, 평가를 받는 수학교사 그리고 수학 수업 관찰횟수는 무한 전집에서 임의로 표집된 것으로 가정하여 임의국면으로 정의하였다. 먼저 IQA 도구를 활용하여, 교사들마다 2번의 수업관찰 평가점수가 주어지는 경우 이를 일반화가능도 이론의 분석 체계로 표현하면, 단변량 $t \times o$ 설계가, 그리고 교사들마다 동일한 문항에 대해 2번의 수업관찰 평가점수가 주어지는 경우, 단변량 $t \times o \times i$ 설계가 된다. 이러한 단변량 일반화가능도 분석의 $t \times o$ 설계와 $t \times o \times i$ 설계의 분산성분을 추정하기 위한 G-연구와, 이를 바탕으로 오차국면의 수를 조정함으로써 효율적인 측정조건을 탐색하기 위한 D-연구를 수행하기 위해서는 GENOVA(Crick & Brennan, 1983) 프로그램을 사용하였다. 다음으로 IQA 도구를 활용하여, 교사들마다 각 채점영역에 따라 2번의 수업관찰 평가점수가 주어지는 경우 이를 다변량 일반화가능도 이론의 분석 체계로 표현하면, $t^{\bullet} \times o^{\bullet}$ 설계가, 그리고 각 채점영역마다 2번의 수업관찰 횟수는 동일하지만, 각 채점영역 별 문항이 다른 경우 이를 다변량 일반화가능도 이론의 분석 체계로 표현하면, $t^{\bullet} \times o^{\bullet} \times i^{\circ}$

설계가 된다. 여기서 닫힌 원(●)은 해당 국면이 고정국면인 채점영역 국면과 교차됨을 의미하며, 열린 원(○)은 해당 국면이 채점영역 국면에 내재됨을 의미한다(Brennan, 2001). 이러한 다변량 일반화가능도 분석의 $t^{\bullet} \times o^{\bullet}$ 설계와 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계의 분산성분과 공분산성분을 추정하기 위한 G-연구와, 이를 바탕으로 오차국면의 수를 조정함으로써 효율적인 측정조건을 탐색하기 위한 D-연구를 수행하기 위해서는 mGENOVA(Brennan, 1991) 프로그램을 사용하였다.

Ⅳ.연구 결과

1. 단변량 $t \times o$ 설계

단변량 단일 국면 교차 $t \times o$ 설계의 각 효과에 해당하는 분산 성분 추정치와 해당 효과가 차지하는 비율은 <표 5>와 같다. 전집점수 분산인 교사 분산이 57.66%로 상대적으로 가장 크게 나타났다. 다음으로 잔차 분산이 38.79%로 크게 나타났으며, 관찰횟수에 의한 분산은 3.55%로 가장 낮게 나타났다. 이는 IQA 도구를 활용하여 관찰된 수업관찰 평가점수에 가장 큰 영향을 미치는 요인이 수학교사의 교수 질 차이이며, 관찰횟수에 따라 교사들의 평가점수의 상대적 순위가 달라지지는 않는다는 것을 의미한다.

G-연구 결과를 바탕으로 수행한 D-연구 결과는 <표 6>과 같다. 원자료에 사용된 IQA 도구를 활용하여 관찰된 수업관찰 평가의 일반화가능도계수는 0.75이며, 일반화가능도계수 0.8에 도달하기 위해서는 관찰횟수가 3번 이상이 되어야 하는 것으로 나타났다. 일반적으로 신뢰도의 크기가 얼마 이상이 되어야 한다는 절대적인 기준은 제시되어 있지 않고, 그 기준 또한 평가도구나 연구자들마다 다르다. 그러나 Shavelson, Baxter, 그리고 Gao(1993)와, Dunbar, Koretz, 그리고 Hoover(1992)는 안정된 신뢰도는 0.80이상을 주장하며, Brennan(2001)과 Webb, Shavelson, 그리고 Maddahian(1983)을 비롯하여 일반화가능도 분석을 적용한 많은 연구들은 일반적으로 일반화가능도계수 0.80이상을 적정수준의 신뢰도로 판단하고 있으므로, 이 연구에서도 0.80을 기준으로 정하였다.

<표 5> $t \times o$ 설계의 G-연구 결과

효과	자유도	제곱합	평균제곱	분산성분 추정치 (%)
교사(t)	95	21.0176	0.2212	0.0828 (57.66)
관찰횟수(o)	1	0.5453	0.5453	0.0051 (3.55)
잔차(to, e)	95	5.2870	0.0557	0.0557 (38.79)
전체	191	26.8499	0.8222	0.1436 (100.00)

<표 6> $t \times o$ 설계의 D-연구 결과

관찰횟수	전집점수분산	상대오차분산	일반화가능도계수
1	0.0828	0.0557	0.5980
2	0.0828	0.0278	0.7485
3	0.0828	0.0186	0.8170

주. 음영부분은 G-연구에서와 같은 표본크기를 나타내는 원자료에 사용된 IQA 도구를 나타냄.

2. 다변량 $t^{\bullet} \times o^{\bullet}$ 설계

다변량 $t^{\bullet} \times o^{\bullet}$ 설계의 각 효과에 해당하는 분산과 공분산 성분 추정치, 그리고 해당 효과가 차지하는 비율은 <표 7>과 같다. 인지적 요구영역과 셋업영역에서는 교사 분산이 각각 49.35% 그리고 59.37%로 가장 크게 나타났으며, 다음으로 잔차, 그리고 관찰횟수 분산이 가장 작게 나타났다. 마찬가지로, 토론영역에서도 관찰횟수 분산이 가장 작게 나타났으며, 잔차 분산이 가장 크게, 다음으로 교사 분산 순으로 나타났다. 일반적으로 잔차 분산이 크게 나타나는 이유는 교사와 관찰횟수의 상호작용효과와 연구 설계에 포함되지 못한 다른 국면들의 효과가 서로 분리되지 못하고, 잔차라는 이름으로 분산 성분에 함께 포함되었기 때문이다(Brennan, 2001; Cronbach et al., 1997, Nuβbaum, 1984; 김성연, 2013, 2014).

G-연구 결과를 바탕으로, 세 평가 영역인 인지적 요구영역, 토론영역, 그리고 셋업영역의 합성점수에 대한 D-연구 결과는 <표 8>와 같다. 이러한 결과를 바탕으로 일반화가능도 이론은 가장 효율적인 측정 조건을 제시할 수 있다. 원자료에 사용된 IQA 도구를 활용하여 관찰된 수업관찰 평가의 일반화가능도계수는 0.72이며, 적정 수준인 일반화가능도계수 0.80에 도달하기 위해서는 각 채점영역의 관찰횟수가 3번 이상이 되어야 하는 것으로 나타났다. 또한 관찰횟수를 원자료와 같은 6번으로 고정한 상태에서 각 채점영역 별 관찰 횟수를 조정하는 $t^{\bullet} \times o^{\circ}$ 설계에서 인지적 요구영역은 1번, 토론영역은 1번, 그리고 셋업영역은 4번으로 증가시킨 D-연구 결과 일반화가능도계수는 0.86까지 증가하였다.

이 설계에서 합성점수는 관찰횟수에 따른 각 채점영역의 평균이 사용되었으므로, 각 채점영역 별로 동일한 가중치를 사용하였다. 이렇게 IQA 도구를 활용하여 산출된 수업관찰 평가의 합성점수는 각 채점영역에서 수학교사의 교수 질이 같은 비율로 반영되어 있음을 전제로 하는데 다변량 일반화가능도 분석은 이 전제가 타당한지에 대한 정보를 제공하고 있다. 수업관찰 평가의 합성점수 차이에 각 채점영역 점수가 같은 비율로 반영되어 있는지를 분석한 결과는 <표 9>에 제시되어 있다. 수업관찰 평가의 합성점수에 토론영역과 셋업영역은 상대적으로 더 많이 반영되었으며, 인지적 요구영역에 대한 정보는 상대적으로 덜 반영되는 것으로 나타났다.

<표 7> $t^{\bullet} \times o^{\bullet}$ 설계의 G-연구 결과

효과	분산성분 추정치 (%)			
	채점영역	인지적 요구	토론	셋업
교사(t)	인지적 요구	0.0944 (49.35)	<0.4573>	<0.5433>
	토론	0.0534	0.1446 (45.77)	<0.5490>
	셋업	0.0621	0.0770	0.1385 (59.37)
관찰횟수(o)	인지적 요구	0.0058 (3.03)		
	토론	0.0021	0.0000 (0.00)	
	셋업	0.0068	0.0023	0.0061 (2.61)
잔차(to, e)	인지적 요구	0.0911 (47.62)		
	토론	0.0578	0.1713 (54.22)	
	셋업	0.0153	0.0461	0.0887(38.02)
분산성분 전체		0.5456 (100.00)	0.5362 (100.00)	1.7675 (100.00)

주1. < >는 측정오차를 고려한 상관계수 값을 표시함.
주2. ( )는 채점영역 별 전체 분산 중 해당 효과의 분산이 차지하는 비율임.

<표 8> $t^{\bullet} \times o^{\bullet}$ 설계의 D-연구 결과

관찰횟수(CD, D, SU)	전집점수분산	상대오차분산	일반화가능도계수
(1, 1, 1)	0.0849	0.0655	0.5645
(2, 2, 2)	0.0849	0.0327	0.7216
(3, 3, 3)	0.0849	0.0218	0.7954
⋮	⋮	⋮	⋮
(I, I, 4)	0.1022	0.0171	0.8564

주. 음영부분은 G-연구에서와 같은 표본크기를 나타내는 원자료에 사용된 IQA 도구를 나타냄.

<표 9> 각 채점영역 점수가 합성점수의 분산에 미친 영향

	인지적 요구	토론	셋업
원자료의 가중치	33.33%	33.33%	33.33%
합성점수 분산에 기여한 비율	27.48%	36.09%	36.44%

3. 단변량 $t \times o \times i$ 설계

단변량 두 국면 완전교차 $t \times o \times i$ 설계의 각 효과에 해당하는 분산 성분 추정치와 해당 효과가 차지하는 비율은 <표 10>과 같다. 문항에 의한 분산은 48.04%로 상대적으로 가장 크게 나타났으며, 관찰횟수에 의한 분산은 0.47%로 상대적으로 작게 나타났다. 이는 IQA 도구를 활용하여 관찰된 수업관찰 평가점수 차이가 관찰횟수의 차이보다는 문항의 특성에 기인하고 있음을 의미한다. 또한 교사와 문항과의 상호작용효과는 15.63%로, 교사와 관찰횟수의 상호작용 효과는 3.19%로 나타났다. 이는 IQA 도구를 활용한 평가점수의 상대적 순위가 관찰횟수의 차이보다는 문항의 특성에 따라 다르게 나타난다고 해석할 수 있다. 전집점수 분산인 교사의 분산은 6.52%로, 잔차 분산은 26.00%로 나타났다.

<표 10> $t \times o \times i$ 설계의 G-연구 결과

효과	자유도	제곱합	평균제곱	분산성분 추정치(%)
교사(t)	95	273.1631	2.8754	0.0699 (6.52)
관찰횟수(o)	1	7.0869	7.0869	0.0050 (0.47)
문항(i)	12	1195.0922	99.5910	0.5147 (48.04)
교사×관찰횟수(to)	95	68.7208	0.7234	0.0342 (3.19)
교사×문항(ti)	1140	699.4463	0.6136	0.1675 (15.63)
관찰횟수×문항(oi)	12	5.1162	0.4264	0.0015 (0.14)
잔차(toi,e)	1140	317.5761	0.2786	0.2786 (26.00)
전체	2495	2566.2015	111.5953	1.0714 (100.00)

G-연구 결과를 바탕으로 수행한 D-연구 결과는 <표 11>과 같다. 원자료에 사용된 IQA 도구를 활용하여 관찰된 수업관찰 평가의 일반화가능도계수는 0.63이며, 문항 수를 4배로 늘려도 0.75에 머물렀다. 적정 수준인 일반화가능도계수 0.80에 도달하기 위해서는 관찰횟수는 3번 이상, 그리고 문항 수는 4배로 늘려야하는 것으로 나타났다. $t \times o$설계와 비교하면 일반화가능도계수가 낮은데, 이는 전집점수분산은 $t \times o$설계에서 더 커지고, 상대

오차분산은 $t \times o$설계에서 더 작아졌기 때문이다. <표 10>에 제시되어 있는 바와 같이 교사와 관찰횟수의 상호작용이 상대오차분산에 포함되어 있기 때문이다.

<표 11> $t \times o \times i$ 설계의 D-연구 결과

관찰횟수	문항	전집점수분산	상대오차분산	일반화가능도계수
1	13	0.0699	0.0685	0.5049
	⋮	⋮	⋮	⋮
	52	0.0699	0.0428	0.6202
2	13	0.0699	0.0407	0.6319
	⋮	⋮	⋮	⋮
	52	0.0699	0.0230	0.7523
3	13	0.0699	0.0314	0.6898
	⋮	⋮	⋮	⋮
	52	0.0699	0.0164	0.8098

주. 음영부분은 G-연구에서와 같은 표본크기를 나타내는 원자료에 사용된 IQA 도구를 나타냄.

4. 다변량 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계

다변량 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계의 각 효과에 해당하는 분산과 공분산 성분 추정치와 해당 효과가 차지하는 비율은 <표 12>와 같다. 인지적 요구영역에서는 문항에 의한 분산과 교사와 문항의 상호작용 효과가 19.69%와 18.57%로 각각 상대적으로 크게 나타났으며, 다음으로 교사에 의한 분산은 14.19%로 나타났다. 관찰횟수에 의한 효과는 1.03%로 상대적으로 작게 나타난 반면에, 교사와 관찰횟수에 의한 상호작용효과는 10.74%로 나타남으로써, 수업관찰 평가점수의 상대적 순위가 관찰횟수의 차이보다는 문항의 특성에 따라 다르게 나타나는 것으로 나타났다. 또한 잔차 분산이 35.56%로 가장 크게 나타났다. 토론영역과 셋업영역에 대해서도 같은 방법으로 해석할 수 있다. 셋업영역에서는 문항에 의한 분산이, 그리고 잔차 분산을 제외하면 토론영역에서는 교사 분산이 가장 크게 나타났다.

<표 12>에 제시되어 있는 상관계수는 수학교사들의 세 영역에 대한 전집점수 사이의 관계를 보여주며, 인지적 요구영역과 토론영역의 상관계수는 0.54로, 인지적요구영역과 셋업영역의 상관계수는 0.75로 그리고 토론영역과 셋업영역의 상관계수는 0.74로 나타났다. 이러한 결과는 한 채점영역에서 높은 점수를 받은 교사들이 다른 채점영역에서도 높은 점수를 받았음을 의미하며, 인지적 요구영역, 토론영역, 그리고 셋업영역으로 구성된 수학교사의 교수 질 개념을 IQA 도구로 측정한 수학 수업관찰 평가가 제대로 측정하고 있다는 것을 나타내는 구인타당도로 설명할 수 있다(Brennan, 2001; Webb et al, 1983).

<표 12> $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계의 G-연구 결과

효과	분산성분 추정치 (%)			
	채점영역	인지적 요구	토론	셋업
t	인지적 요구	0.0774 (14.19)	<0.5403>	<0.7481>
	토론	0.0533	0.1258 (23.46)	<0.7351>
	셋업	0.0620	0.0777	0.0888 (5.02)
o	인지적 요구	0.0056 (1.03)		
	토론	0.0021	0.0001 (0.02)	
	셋업	0.0068	0.0023	0.0054 (0.31)
i	인지적 요구	0.1074 (19.69)		
	토론		0.0927 (17.29)	
	셋업			1.0227 (57.86)
to	인지적 요구	0.0586 (10.74)		
	토론	0.0578	0.0624 (11.64)	
	셋업	0.0153	0.0461	0.0113 (0.64)
ti	인지적 요구	0.1013 (18.57)		
	토론		0.0375 (7.00)	
	셋업			0.2487 (14.07)
oi	인지적 요구	0.0013 (0.24)		
	토론		0.0000 (0.00)	
	셋업			0.0039 (0.22)
toi,e	인지적 요구	0.1940 (35.56)		
	토론		0.2177 (40.60)	
	셋업			0.3867 (21.88)
분산성분 전체		0.5456 (100.00)	0.5362 (100.00)	1.7675 (100.00)

주1. < >는 측정오차를 고려한 상관계수 값을 표시함.
주2. ( )는 채점영역 별 전체 분산 중 해당 효과의 분산이 차지하는 비율임.

G-연구 결과를 바탕으로, 인지적 요구영역, 토론영역, 그리고 셋업영역의 합성점수에 대한 D-연구 결과는 <표 13>과 같다. 이러한 결과를 바탕으로 일반화가능도 이론은 가장 효율적인 측정 조건을 제시할 수 있다. 원자료에 사용된 IQA 도구를 활용하여 관찰된 수업관찰 평가의 일반화가능도계수는 0.65이며, 적정수준인 일반화가능도계수 0.80에 도달하기 위해서는 각 채점영역의 관찰횟수가 8번 이상이 되어야 하는 것으로 나타났다. 관찰횟수를 각 채점영역별로 2번으로 고정하고, 원자료와 같은 문항의 총 수인 13개로 고정한 상태에서, 채점영역 내의 문항 수를 인지적 요구영역은 1개, 토론영역은 11개 그리고 셋업영역은 1개로 조정하면, 일반화가능도계수는 0.72까지 증가하였다. 또한 관찰횟수를 원자료와 같은 6번으로 고정한 상태에서, 각 채점영역 별 관찰횟수와 채점영역 내 문항 수를 조정하는 $t^{\bullet} \times o^{\circ} \times i^{\circ}$ 설계에서 인지적 요구영역은 1번, 토론영역은 4번, 그리고 셋업영역은 1번, 그리고 인지적 요구영역 내 문항은 1개, 토론영역 내 문항은 6개, 그리고 셋업영역 내 문항은 1개로 조정하였을 때, 일반화가능도계수는 0.81까지 증가하였다.

<표 13> $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계의 D-연구 결과

관찰횟수(CD, D, SU)	문항(CD, D, SU)	전집점수분산	상대오차분산	일반화가능도계수
(1, 1, 1)	(6, 2, 5)	0.0714	0.0670	0.5157
	(1, 11, 1)	0.1089	0.0809	0.5736
(2, 2. 2)	(6, 2, 5)	0.0714	0.0392	0.6454
	(1, 11, 1)	0.1089	0.0427	0.7182
⋮	⋮	⋮	⋮	⋮
(8, 8, 8)	(6, 2, 5)	0.0714	0.0184	0.7954
⋮	⋮			
(1, 4, 1)	*(1, 6, 1)*	0.1169	0.0278	0.8078

주. 음영부분은 G-연구에서와 같은 표본크기를 나타내는 원자료에 사용된 IQA 도구를 나타냄.

이 설계에서 합성점수는 각 채점영역 내의 문항 수들이 반영되었으므로, 각 채점영역 별로 문항 수에 대한 가중치를 사용하였다. 이렇게 IQA 도구를 활용하여 산출된 수업관찰 평가의 합성점수는 각 채점영역에서 수학교사의 교수 질이 문항 수와 비례하는 비율로 반영되어 있음을 전제로 하는데, 다변량 일반화가능도 분석은 이 전제가 타당한지에 대한 정보를 제공하고 있다. 수업관찰 평가의 합성점수 차이에 각 채점영역 점수가 문항 수와 비례하게 반영되어 있는지를 분석한 결과는 <표 14>에 제시되어 있다. 수업관찰 평가의 합성점수에 토론영역은 원자료에서 의도한 바와 비슷하게, 셋업영역은 상대적으로 더 많이 반영되었으며, 인지적 요구영역에 대한 정보는 상대적으로 덜 반영되는 것으로 나타났다.

<표 14> 각 채점영역 점수가 합성점수의 분산에 미친 영향

	인지적 요구	토론	셋업
원자료의 가중치	46.15%	15.39%	38.46%
합성점수 분산에 기여한 비율	43.81%	15.92%	40.27%

5. 연구 설계 비교

연구 설계들의 비교를 용이하게 하기 위해, 원자료와 같은 표본크기를 갖는 D-연구결과들을 중심으로 재정리한 결과는 <표 15>-<표 17>과 같다. <표 15>에 제시된 바와 같이, 국면이 늘어날수록 대체적으로 일반화가능도계수는 감소하는 것으로 나타났다.

<표 15> 4가지 설계의 D-연구 결과

	$t \times o$	$t^{\bullet} \times o^{\bullet}$	$t \times o \times i$	$t^{\bullet} \times o^{\bullet} \times i^{\circ}$
(합성점수)전집점수	0.0828	0.0849	0.0699	0.0714
(합성점수)상대오차	0.0278	0.0327	0.0407	0.0392
(합성점수)일반화가능도계수	0.7485	0.7216	0.6319	0.6454

주. (합성점수)는 영역 내 문항의 가중치를 반영한 $t^{\bullet} \times o^{\bullet}$ 그리고 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계에 해당함.

<표 16>은 국면이 가장 많이 포함됨으로써, 이론적으로 가장 적합한 연구설계인 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 를 기준으로 하여, 다른 연구설계 결과들의 편의를 나타내고 있다. 가장 큰 편의는 문항효과와 채점영역 효과를 모두 무시

하였을 때, 0.1031정도 과대추정 되었다. 다음으로 문항효과만을 무시하였을 때, 그리고 채점영역 효과만을 무시하였을 때의 순으로 나타났다.

<표 16> D-연구결과의 편의

	$t \times o$	$t^{\bullet} \times o^{\bullet}$	$t \times o \times i$
(합성점수)전집점수	0.0114	0.0135	-0.0015
(합성점수)상대오차	-0.0114	-0.0065	0.0015
(합성점수)일반화가능도계수	0.1031	0.0762	-0.0135

주. $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계를 기준으로 함.

구체적인 문항효과와 채점영역 효과는 <표 17>에 제시하였다. 각 효과는 연구 설계 중 오직 문항 국면만 무시되었을 때, 그리고 채점영역 국면만 무시되었을 때의 차로 계산하였다. 문항 효과가 채점영역 효과보다 전반적으로 크다는 것을 알 수 있다. 문항 국면을 고려하지 않은 경우, IQA 도구를 활용하여 관찰된 수업관찰 평가의 신뢰도는 과대추정 되는 것으로 나타났다. 이는 $t \times o$ 설계에서는 잔차 분산으로 포함되었던 교사와 문항과의 상호작용 효과가 $t \times o \times i$ 설계에서는 상대오차분산에 포함되기 때문이다. 또한 채점영역 국면을 고려하는 경우, 일반화가능도계수의 차이는 0.03 또는 0.01로 작은데, 이는 채점영역간의 상관계수가 높음을 이유로 들 수 있다. 채점영역 간의 상관계수가 높을 때, 채점영역 내 분산의 평균은 채점영역 간 공분산의 평균보다 훨씬 크지 않을 수 있기 때문이다(Brennan, 2001; Li & Brennan, 2007).

<표 17> 문항과 채점영역 효과

	문항 효과		채점영역 효과	
	$t \times o - t \times o \times i$	$t^{\bullet} \times o^{\bullet} - t^{\bullet} \times o^{\bullet} \times i^{\circ}$	$t \times o - t^{\bullet} \times o^{\bullet}$	$t \times o \times i - t^{\bullet} \times o^{\bullet} \times i^{\circ}$
전집점수	0.0129	0.0135	-0.0021	-0.0015
상대오차	-0.0129	-0.0065	-0.0049	0.0015
일반화가능도계수	0.1166	0.0762	0.0269	-0.0135

Ⅴ. 결론 및 논의

이 연구는 수학교사의 수업을 관찰함으로써 수학 교수 질을 평가하는 IQA 도구를 활용하여 각 측정상황에서 발생하는 오차요인들의 상대적인 영향력을 살펴보고, 이를 바탕으로 효율적인 최적의 측정조건을 탐색하는 방법을 제시하였다. 연구 자료는 2007년부터 NSF의 지원으로 수행되고 있는 MIST 프로젝트의 종단 자료 중, 이틀 연속 인지적 요구영역, 토론영역 그리고 셋업영역에 평가점수가 부여된 3차년도와 4차년도 수학교사들의 자료를 활용하여, 단변량과 다변량 일반화가능도 분석을 수행하였다. 연구 결과를 바탕으로 일반화가능도 분석이 어떻게 우리나라 수학교사의 수업관찰을 바탕으로 교수 질을 평가하는 도구를 개발하고, 향상시키는데 기여할 수 있는지에 대해 논하면 다음과 같다.

G-연구 결과, $t \times o$ 설계와 $t^{\bullet} \times o^{\bullet}$에서는 측정의 대상인 교사 분산이 대체적으로 가장 크게 나타났으며, 관찰횟수 분산은 가장 작게 나타났다. 이처럼 교사 분산이 상대적으로 크게 나타났다는 것은 IQA 도구를 활용하여 관찰된 수업관찰 평가점수가 수학 교수 질 차이를 반영하고 있음을 보여주며, 관찰횟수 분산이 작게 나타났다는 것은 관찰횟수에 따라 교사의 수업관찰 평가점수가 달라지지 않고 있음을 보여주고 있다. 이러한 결과는 IQA 도구를 활용하여 관찰된 수학 수업관찰 평가에 관찰횟수만을 국면으로 고려하여 단변량 일반화가능도 분석

을 적용한 Matsumura 외(2008)의 연구와 일치하였다. 그러나 문항국면을 함께 고려하는 $t \times o \times i$ 설계와 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$에서는 대부분 문항 분산이 가장 크게 나타났으며, 다음으로 교사 분산, 그리고 관찰횟수 분산은 마찬가지로 가장 작게 나타났다. 이러한 결과는 미국의 BTAP를 위한 채점자 훈련과정에 참가한 교사평가 자료에 단변량 일반화가능도 분석을 적용한 김성숙(1989a, 1989b, 1992)의 연구와 교원능력개발평가에 사용되는 수업관찰 평가 자료에 다변량 일반화가능도 분석을 적용한 김성연(2014a)의 연구와 부분적으로 일치하였다. 또한 각 채점영역 간의 공분산과 측정오차를 고려한 상관계수를 살펴보면, 한 채점영역에서 높은 점수를 받은 교사들은 다른 영역에서도 높은 점수를 받았다는 것을 알 수 있다.

이처럼, 설계에서 고려하는 국면에 따라 그 국면들의 상대적인 분산의 크기가 달라진다는 것은, 일반적으로 연구자들이 수업관찰 평가도구에 가장 많이 사용하는 채점자간 신뢰도나 Cronbach α 사용의 한계점을 지적할 수 있다. 즉 이러한 두 신뢰도를 일반화가능도이론 입장에서 설명하면, 채점자간 신뢰도는 오직 채점자 국면만을, Cronbach α는 오직 문항 또는 시기(occasion)과 같은 단일 국면만을 설계에 포함하게 된다. 따라서 이 외의 다른 국면들이 설계에 포함되는 경우 고전검사이론의 진점수 분산에 해당하는 측정의 대상인 교사의 분산이 가장 크게 나타났다하더라도, 다른 국면이 포함됨에 따라 다시 재조정됨으로써 이 부분이 오차점수 분산에 기여할 수 있다는 것을 의미한다.

우리나라에서도 교원능력개발평가에 수업관찰이 포함되어 있지만, 아직 초기단계로 구체적인 평가틀이 제공되지 못하고 있는 것이 현실이다. 예를 들어, 교원능력개발평가는 9월부터 11월, 근무성적평정은 12월 말, 그리고 교원상여금은 2월로, 각각 유사한 내용으로 평가가 따로 실시되어 현장 교원의 피로감이 증가된다는 지적을 반영하여, 2014년 2월 '2단계 교원평가제도 개선방안'의 하나로 2014년 하반기 전국 67개 초중고를 연구시범학교로 지정해 교원능력개발평가, 근무성적평정, 그리고 근무성적평정의 지표를 조정하고, 이를 같은 기간 안에 실시하겠다고 발표하였다. 그러나 여전히 교장과 교감은 인사평정자라는 점에서 채점자에 포함시켜야 하는지, 포함된다면 동료교원을 포함하여 몇 번을 관찰하여야 하는지, 시간은 어느 정도나 관찰해야 하는지에 대한 논의가 계속되고 있다(최대현, 2014). 따라서 김성연(2014)에서 제시한 것처럼, 연구시범학교에서 수집한 교원능력개발평가 자료에 일반화가능도 분석의 G-연구를 수행하는 것이 필요하다. 이러한 자료를 분석하여 얻어지는 단변량 일반화가능도 G-연구의 분산성분, 또는 다변량 일반화가능도 G-연구의 분산과 공분산 성분을 바탕으로 수업관찰 평가의 오차 국면의 상대적인 중요성을 파악할 수 있다. 예를 들어, 평가 횟수, 채점자, 채점자 특성, 평가 문항 수, 또는 평가 문항 별 점수 등의 상대적인 분산 성분의 크기를 살펴봄으로써, 어떤 오차 국면이 측정의 일반화과정을 저해하는 지 설명하고, 각 오차 국면들을 동시에 파악할 수 있는 분석 틀을 제공함으로써 보다 객관적이고 안정적인 자료를 수집하기를 기대할 수 있다.

D-연구 결과, 적정수준의 일반화가능도 계수인 0.80을 확보하기 위한 최적의 측정 조건을 살펴보면, $t \times o$와 $t^{\bullet} \times o^{\bullet}$설계의 경우 3번 이상의 관찰횟수, $t \times o \times i$ 설계의 경우 관찰횟수는 3번 이상, 문항 수는 4배 이상, 그리고 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계의 경우 문항 수가 원자료와 같은 경우 8번 이상의 관찰횟수가 필요한 것으로 나타났다. 또한 원자료와 같은 총 문항 수를 기준으로 했을 때는, $t^{\bullet} \times o^{\bullet}$설계의 경우 인지적 요구영역은 1개, 토론영역은 1개, 그리고 셋업영역은 4개로 구성하는 경우 원자료 기준으로 0.72였던 일반화가능도계수는 0.86으로 증가하였다. 또한 $t^{\bullet} \times o^{\bullet} \times i^{\circ}$ 설계의 경우 관찰횟수는 인지적 요구영역에서 1번, 토론영역에서 4번, 그리고 셋업영역에서 1번으로, 문항 수는 인지적 요구영역에서 1개, 토론영역에서 6개, 그리고 셋업영역에서 1개로 구성하는 경우, 원자료 기준으로 0.65였던 일반화가능도계수는 0.81까지 증가하였다. 이처럼 D-연구 결과는 평가를 계획하는 의사결정자가 설정한 적정 수준의 신뢰도에 도달하는데 있어, 평가 상황에 따라 실질적으로 의사결정에 도움을 주는 합리적인 측정 조건을 제시할 수 있다. 우리나라에서도 교원능력개발평가의 공정성과 객관성을 확보하기 위해, 2010년 동료교사 지표별로 2문항 이상씩 총 36문항, 학생 및 학부모의 경우 지표별로 1

문항 이상씩 18문항 이상이었던 것을 2011년 동료교사의 경우 13문항으로, 학생은 5문항으로, 학부모의 경우는 2문항에서 5문항으로, 그리고 2012년 동교교사의 경우 13문항 이상으로, 학생과 학부모는 모두 5문항 이상으로, 연도에 따라 때로는 문항 수를 축소하고 확대해가면서 수정을 거듭하고 있다.

또한 평가문항의 경우 2010년은 담임, 교과(전담)교사, 비교과교사를 일부 차별화했던 것에서, 2011년에는 담임, 교과(전담)교사, 비교과교사를 모두 차별화, 그리고 2012년에는 담임, 교과(전담)교사, 비교과교사를 모두 차별화할 뿐만 아니라, 담당교과별 문항도 차별화 하는 등 해마다 개선된 내용들(교육과학기술부, 2013)이 반영되고 있다. 그러나 여전히 평가지표에 대해서는 교과(전담)교사의 경우 학습지도 영역의 문항을 강화하여 학습지도 9문항, 생활지도영역 4문항으로, 그리고 담임교사의 경우 생활지도영역의 문항을 강화하여 학습지도 8문항, 생활지도영역 5문항으로 13문항 이상을 전국공통기준으로 정하고 있을 뿐, 이 두 영역의 평가지표를 한 개의 지표로 합할지, 아니면 그대로 두고 반영 비율 등을 조정할지 등이 확정되지 못한 상황이다.

최근 교육과학기술부가 밝힌 일정을 살펴보면, 6월 연구시범학교에 적용할 평가지표를 개선하는 운영안을 만들고, 8월 확정해 하반기에 운영할 예정이며, 이 결과를 바탕으로 2015년 3월에 평가의 개선안을 마련해, 9월에 전면 실시한다는 계획을 발표하였다(최대현, 2014). 따라서 김성연(2014)에서 제시한 바와 같이, 연구학교에 대해 시범적용해서 수집한 교원능력개발평가 자료에 일반화가능도 이론의 D-연구를 수행하는 것이 필요하다. 이를 통해 일반화가능도 이론의 G-연구 결과를 바탕으로 얻어지는 D-연구 결과를 통해, 실제 사용되는 측정 절차의 각 조건을 효율적으로 최소화하면서 적정 수준 이상의 신뢰도 계수를 산출할 수 있는 최적의 측정 구조를 탐색하기를 기대할 수 있다. 즉 과목의 특성을 고려하여 수학교사의 경우 채점영역 별로 관찰횟수와 문항 수를 조절함으로써, 채점자들의 채점에 따른 수고를 줄여주면서, 동시에 적정 수준의 신뢰도를 얻을 수 있다.

또한 다변량 일반화가능도 분석을 통한 연구결과는 평가도구를 개발하는 연구자들이 현실에서 당연시 여기는 점수산출 방법에 대한 가정이 틀릴 수 있다는 점을 보여주었다. 일반적으로 평가도구에 대한 합성점수는 각 채점영역을 더하거나 평균을 이용하는 경우가 많은데, 이것은 각 채점영역의 가중치가 문항 수나 또는 같은 비율로 반영되어 있다고 가정하는 것이다. 그러나 이 연구의 결과가 제시하는 것처럼, 각 채점영역의 분산이나 각 채점영역 간의 공분산에 따라 반영되는 비율이 달라짐으로써 실질 가중치는 달라지게 된다. 따라서 연구자가 상정한 처음의 가중치를 얻었는지에 대한 검증이 필요하다. 또한 채점자의 평가기준은 지침과 훈련과정을 통해 객관화 할 수 있으므로, 일반화가능도 분석을 수행하여 몇 번을 평가하여야 하는지, 몇 개의 문항을 사용하여야 하는지, 그리고 얼마동안을 관찰하여야 하는지와 같은 효율적인 측정조건을 바탕으로 후속 측정과정을 설계할 수 있다.

지금까지의 연구결과가 교육제도, 교육환경, 문화 그리고 구성원 등 모든 면들이 다른 미국의 IQA 도구를 우리나라 수학수업 관찰을 평가하는 도구로 바로 적용하는 데는 한계가 있을 것이다. 그러나 다양한 오차원을 고려함에 따라 실용성을 고려한 검사의 측정학적 특성을 탐색함으로써 효율적인 측정조건을 확정하기 위한 방법론적인 틀로서 기능하여 수학교사의 수업관찰을 바탕으로 교수 질 평가 시 측정학적으로 더욱 바람직한 의사결정이 이루어 질 수 있는 기반을 제공할 것으로 기대된다. 마지막으로 이 연구의 제한점과 후속 연구를 제안하면 다음과 같다. 첫째, 관찰횟수와 문항 외에 채점자를 비롯하여 수업 내용 등을 임의국면으로 설정하여 분석할 필요가 있다. 이 연구에서는 이미 IQA 도구로 측정된 미국 수학교사의 수업관찰 자료를 활용하여 채점영역은 고정 국면으로, 수업관찰횟수와 문항은 임의국면으로 다룬 설계들에 초점을 맞추었다. 따라서 후속연구에서는 MET에서 사용된 설계처럼, 교사들의 수업이 관찰된 시점에 따라 수업 내용이 다를 수 있음을 고려하여, 우리나라 수학 교사를 대상으로 수업 내용 국면이 측정의 대상인 교사에 내재되어 있으며, 채점자를 함께 국면으로 고려할 것을 제안한다. 둘째, 이 연구에서는 수학 교수 질 평가를 위해 IQA 도구를 사용하였지만, 이 외에도 수학 수업 관찰을 기반으로 하는 다양한 평가도구들이 개발되어 있다. 따라서 교사들을 대상으로 수학 교수에 대한 지식(Mathematical Knowledge for teaching, MKT) 도구(Hill et al., 2004), 수학 교수 질(Mathematical Quality

of Instruction, MQI) 도구(Hill et al., 2008), 그리고 질 높은 수학 교수에 대한 비전(Vision of High Quality Mathematical Instruction, VHQMI) 도구(Munter & Correnti, 2011)를 활용하여 일반화가능도 분석을 수행할 것을 제안한다. 마지막으로 수업관찰 평가의 최종점수를 산출할 때 일반적으로 각 채점영역 점수를 더하거나 평균을 냄으로써, 각 채점영역 가중치는 문항 수에 비례하거나 같은 비율을 사용하게 된다. 그러나 다변량 일반화가능도 이론을 적용한 분석은 각 영역 점수의 분산이나 상관관계에 따라 채점영역별 최적의 가중치를 산출해 준다. 따라서 합성점수에 대한 분산과 공분산을 바탕으로 채점영역 별 최적의 상대 가중치를 탐색하기를 제안한다.

참 고 문 헌

교수학습개발센터 (2014). 교원전문성. 서울: 한국교육과정평가원.

교육과학기술부 (2013). 내실 있는 운영을 위한 2013년 교원능력개발평가제 개선 방안.

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.

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

김성숙 (1989a). 일반화가능도 이론을 이용한 교사행위 관찰에 있어서 오차원 분석. 교육평가연구, **3(1)**, 211-219.

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

김성숙 (2001). 일반화가능도이론. 서울: 교육과학사.

김성숙 (2006). e-learning 강의평가 도구의 일반화가능도와 평가활용의 최적화 조건. 교육평가연구, **19(1)**, 305-322.

김성연 (2014a). 미국의 수업관찰평가 분석을 통한 우리나라 교원능력개발평가에서의 다변량 일반화가능도 이론 활용성 탐색. 한국교육, **41(1)**, 5-29.

김성연 (2014b). 미국 테네시 주 벤더빌트대학교 영재교육센터 프로그램이 우리나라 영재교육에 주는 시사점 탐색. 영재교육연구, **24(2)**, 217-243.

김성연・한기순 (2013). 관찰・추천제에 의한 수학영재 선발 시 사용되는 교사추천서와 자기소개서 평가에 대한 다변량 일반화가능도 이론의 활용. 영재교육연구, **23(5)**, 671-698.

김성연・한기순 (2014). 수학영재 선발에서 교사추천서와 자기소개서 채점내용 가중치에 따른 신뢰도 분석. 영재와 영재교육, **13(1)**, 43-65.

김성찬・김성연・한기순 (2012). 관찰・추천에 의한 수학영재 선발 시 사용되는 교사추천서와 자기소개서 평가에 대한 일반화가능도 이론의 활용. 한국수학교육학회 시리즈 E <수학교육 논문집>, **26(3)**, 251-271.

김현철 (2003). 일반화가능도 이론에 대한 대학평가의 신뢰도 추정과 효율적인 평가설계의 탐색. 교육학연구, **41(4)**, 49-70.

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

이대현・최승현 (2006). 수학과 좋은 수업 사례에 대한 질적분석. 한국학교수학회논문집, **9(3)**, 249-263.

이향 (2012). 말하기 수행 평가에서 발음 범주 채점의 최적화 방안 연구-일반화가능도 이론을 활용하여. 한국어교육, **23(2)**, 301-329.

임형·김성숙 (2005). 임상수행능력평가의 오차요인 탐색과 신뢰도 연구. 교육평가연구, **18(1)**, 27-46.

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

임찬빈·이화진·곽영순·강대현·박영석 (2004). 수업평가 기준 개발 연구(1): 일반 기준 및 교과(사회, 과학, 영어). 서울: 한국교육과정평가원.

조재윤 (2009). 일반화가능도 이론을 이용한 쓰기 평가의 오차원 분석 및 신뢰도 추정 연구. 국어교육, **128**, 325-357.

최대현 (2014). 교원평가·근평·성과금 일원화 물밑작업-교육부, 67개 시범학교에서 3개 평가 지표 조정 적용키로. 교육희망, http://news.eduhope.net/sub_read.html?uid=15896

최승현·임찬빈 (2006). 수업평가 매뉴얼-수학과 수업평가 기준. 서울: 한국교육과정평가원.

최숙기 (2012). 국어과 수행 평가의 평가자 신뢰도 보고 방안 탐색-고등학생 요약문 평가 결과를 중심으로. 작문연구, **14**, 395-424.

한혜정 (2012). 미국 뉴욕 주 교원임용 중등 교수역량 지필평가가 우리나라 교원임용시험에 주는 시사점 탐색. 한국교육, **39(2)**, 129-155

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal,* **113(1)**, 76-104.

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

Brennan, R. L. (1991). Manual for mGENOVA. A computer program for computing variance-covariance component. Iowa City, IA: Iowa Testing Program, University of Iowa.

Choi, Sungsook Kim. (1989b). *An analysis of sources of variation in teacher behaviors using generalizability thoery*. Doctoral dissertation. University of Virginia, Charlottesville, VA.

Cobb, P., & Jackson, K. (2011). Towards an empirically grounded theory of action for improving the quality of mathematics teaching at scale. *Mathematics Teacher Education and Development,* **13(1)**, 6-33.

Colorado Department of Education. (2012). 191 into action. Retrieved from http://www.coloradoea.org/

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A GENeralized Analysis Of VAriance System*, ACT Technical Bulletin, 43, The American College Testing Program.

Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement,* **57(3)**, 373-399.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied measurement in education,* **4(4)**, 289-303.

Eckert, J. M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay?. *Phi Delta Kappan,* **91(8)**, 88-92.

Gläser-Zikuda, M., & Fuß, S. (2008). Impact of teacher competencies on student emotions: A multi-method approach. *International Journal of Educational Research,* **47(2)**, 136-147.

Hattie, J. (2003). *New Zealand education snapshot: with specific reference to the Yrs 1-13 years.* Knowledge Wave Trust.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. *Second Handbook of Research on Mathematics Teaching and Learning,* **1**, 371-404.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction,* **26(4)**, 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher,* **41(2)**, 56-64.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal,* **105(1)**, 11-30.

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. research paper. MET Project. *Bill & Melinda Gates Foundation.*

Jackson, K., Garrison, A., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education,* **44(4)**, 646-682.

Jacobs, H. (2010). Race to the top. *EMBO reports,* **11(2)**, 73-73.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation.*

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: combining high-quality observations with student surveys and achievement gains. MET Project. *Bill & Melinda Gates Foundation.*

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher,* **39(8)**, 591-598.

Kyriakides, L., & Creemers, B. P. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education,* **34(5)**, 521-545.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education,* **14**, 25-47.

Li, D., & Brennan, R. (2007). A multi-group generalizability analysis of a large-scale reading comprehension test. *In annual meeting of the National Council on Measurement in Education. Chicago, IL.*

Louisiana Act 54. (2010). Louisiana Departmemt of Education. Retrieved from http://www.louisianabelieves.com

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the instructional quality assessment. CSE Technical Report 681. *National Center for Research on Evaluation, Standards, and Student Testing* (CRESST).

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment,* **13(4)**, 267-300.

Measures of Effective Teaching. (2009). Measures of effective teaching. Retrieved from www.metproject.org

Middle-school Mathematics and the Institutional Setting of Teaching Project. (2007). MIST Project. Retrieved from http://peabody.vanderbilt.edu/departments/tl/teaching_and_learning_research/mist/index.php

Munter, C., & Correnti, R. (2011). *Developing visions of high-quality mathematics instruction.* Paper presented at the National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (1991, 1993). *Principles and standards for school mathematics.* Reston, VA: Author.

National Center for Teacher Effectiveness. (2011). *Online poll of states engaged in reform of teacher evaluation system.* Cambridge, MA: Authors.

Nußbaum, A. (1984). Multivariate generalizability theory in educational measurement: An empirical study. *Applied Psychological Measurement,* **8(2)**, 219-230.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational evaluation and policy analysis,* **26(3)**, 237-257.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review,* **94(2)**, 247-252.

Rowan, B., Correnti, R., & Miller, R. (2002). What large-scale survey research tells us about teacher effects on student achievement: insights from the prospects study of elementary schools. *The Teachers College Record,* **104(8)**, 1525-1567.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Oxford, UK: Pergamon.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement,* **30(3)**, 215-232.

Teddlie, C., & Reynolds, D. (Eds.). (2000). *The international handbook of school effectiveness research.* London: Falmer Press.

Tennessee Department of Education. (2014). *Tennessee Educator Acceleration Model.* Retrieved from team-tn.org

Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Pyans, Jr.(Ed.), *Genrealizability theory: Inferences and practical applications* (pp67-81). San Francisco, CA: Jossey-Bass.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project.*

Exploring the Application of Generalizability Theory to Mathematics Teacher Evaluation for Professional Development in Korea Based on the Analysis of Instructional Quality Assessment of Mathematics Teachers in the U.S.

Kim Sungyeun
Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.
E-mail : sykim0401@snu.ac.kr

The purpose of this study was to suggest methods to apply generalizability theory to mathematics teacher evaluation using classroom observations in Korea by analysing mathematics teachers in the U.S. using the instructional quality of assessment instrument as an illustrative example. The subjects were 96 teachers participating in Year 3 and Year 4 from the Middle-school Mathematics and the Institutional Setting of Teaching (MIST) project funded by the National Science Foundation since 2007. The MIST project investigates the following question: What does it takes to support mathematics teachers' development of ambitious and equitable instructional practices on a large scale (MIST, 2007). This study examined data based on both the univariate generalizability analysis using GENOVA program and the multivariate generalizability analysis using mGENOVA program. Specifically, this study determined the relative effects of each error source and investigated optimal measuring conditions to obtain the suitable generalizability coefficients. The methodology applied in this study can be utilized to find effective optimal measurement conditions for the mathematics teacher evaluation for professional development in Korea. Finally, this study discussed limitations of the results and suggested directions for future research.

* ZDM Classification : C73
* 2000 Mathematics Subject Classification : 97C40
* Key Words : mathematics instructional quality assessment, multivariate generalizability analysis, univariate generalizability analysis
* The data comes from a larger study, supported by the National Science Foundation under grants No.ESI-0554535 and No.DRLL-1119112. I thank Paul Cobb, Tom Smith, Erin Henrick, Anne Wilhelm, and other members of the research team for the MIST project from which this data comes.