ORIGINAL ARTICLE

The School Effect on the Reliability of Clinical Performance Examination in Medical Schools

Mi Kyoung Yim[1] and Gue Min Lee[2]

[1]National Health Personnel Licensing Examination Board, and [2]Department of Education, Yonsei University, Seoul, Korea

임상수행능력평가(CPX)에서 학교효과가 신뢰도에 미치는 영향

[1]한국보건의료인국가시험원, [2]연세대학교 교육과학대학 교육학부

임미경[1], 이규민[2]

Purpose: The purpose of this study is to test the reliability of the clinical performance examination (CPX) using Generalizability theory (G-theory). Through G-theory, the effects of not only students and tasks but also the school will be analyzed as primary sources of error, which can affect the interpretation of the reliability of the CPX.

Methods: One thousand three hundred nineteen students from 16 medical schools that participated in the Seoul-Gyeonggi CPX Consortium 2008 were enrolled. In our research design, we suppose that student is nested within school and crossed with task. Data analysis was conducted with urGenova.

Results: According to our analysis, the percentage of error variance was 6.2% for school, 14.9% for student nested within school, 14.4% for task, and 3% for interaction between school and task. An effect of school on students was observed, but the interaction between task and school was insignificant. When student is nested within school, the universe score decreased and the g-coefficient was less than the g-coefficient of the p × t (p: studentm, t: task) design.

Conclusion: The results show that generalizability theory is useful in detecting various error components in the CPX. Using the generalizability theory to improve the technical quality of performance assessments provides us with greater information compared with traditional test theories.

Key Words: Clinical performance examination, Generalizability theory, Reliability

서론

의사 국가시험에 실기시험 도입이 본격화되면서 의과대학 교육과정에서 임상수행능력평가(clinical performance examination, CPX)가 활성화되고 있다. CPX는 의과대학 학생 평가에 수행평가방법을 적용한 것인데, 수행평가란 학습자가 습득한 지식, 기능이나 기술을 실제 생활이나 인위적 평가 상

Received: May 11, 2010 • Revised: July 21, 2010 • Accepted: August 11, 2010
Corresponding Author: Mi Kyoung Yim
Department of Research and Development, National Health Personnel Licensing Examination Board, 679-30 Jayang-dong, Gwangjin-gu, Seoul 143-873, Korea
Tel: +82.2.2087.8852 Fax: +82.2.2087.8885 email: mkyim@kuksiwon.or.kr

Korean J Med Educ 2010 Sep; 22(3): 215-223.
doi: 10.3946/kjme.2010.22.3.215.
pISSN: 2005-727X eISSN: 2005-7288

황에서 얼마나 잘 수행하는지 혹은 어떻게 수행할 것인지를 서술, 관찰, 면접 등의 다양한 방법을 통해 종합적이며 전문적으로 판단하는 평가 방법이다[1]. 따라서 의사의 진료수행능력을 평가하기 위하여 '표준화환자'라는 연기자가 실제 환자와 같은 상황을 연기하면서 환자를 다루는 학생의 진료 수행능력을 관찰하고 채점하여 점수화한다. 이와 같은 평가방법은 많은 인력과 비용 등 조직적인 시행준비를 요하므로, 의과대학들은 효율적인 과제 개발 및 연기자 교육, 시험 시행을 위하여 CPX 컨소시엄을 구성하고 상호 교류를 활성화하고 있으며 점차 이러한 컨소시엄의 역할이 커지고 있다. 2004년에 8개 의과대학이 형성한 서울-경기 컨소시엄을 시초로, 현재는 지역권역별로 컨소시엄이 존재한다. 컨소시엄을 통해 대학들은 더 많은 과제와 표준화환자 관련 네트워크를 공유하고, 더 많은 수의 학생들에게 적용할 수 있게 되었다[2,3]. 컨소시엄에 속한 의과대학들은 서로 공통된 과제를 사용하거나 표준화환자를 공유하여 시험을 시행하므로 학교별 결과자료는 컨소시엄 데이터로 모아져서 연구자들이 해당 학교의 자료뿐만 아니라 전체의 자료를 분석하고 평가할 수 있게 한다. CPX 점수 분석에 가장 기본적으로 과제의 신뢰도를 보고하는데, 아직까지 CPX의 신뢰도 분석은 Cronbach-α에 의존한 단일 오차 분석이 주를 이루고 있다.

CPX는 관찰에 의해 점수가 부여되는 평가방식일 뿐만 아니라 측정환경에서 다양한 요인이 개입되기 때문에 검사 점수에 영향을 주는 오차의 종류가 다양하다. 단일한 오차원(sources of error, 오차 요인)을 가정하는 고전검사이론에 비해 일반화가능도이론(generalizability theory, G-theory)은 다양한 오차 요인을 고려할 수 있는 측정 이론으로 신뢰도 분석에 유용한 방법을 제공한다. 일반화가능도이론은 고전검사이론을 확대하여 중다오차원을 동시에 고려하는 측정모형에 분산분석(ANOVA)체계를 적용한 이론으로서, 문항이라는 단일한 측정 조건을 갖는 지필 시험과 달리 평가자, 시행 시기, 시행 횟수 등 다양한 측정상황이 동반되기 마련인 수행평가의 점수 분석에 적합한 이론으로 평가받는다[4,5,6].

수행평가 결과에 대하여 측정오차에 대한 다차원적 접근방법을 통해 검사의 신뢰도를 평가하는 연구는 교육학 분야에서 활발히 이루어지고 있으며[7,8,9], 의학교육 분야에서는 아직 제한적이나 점차 증가하고 있다[10,11,12,13].

본 연구에서는 컨소시엄의 CPX 시행 결과에 일반화가능도이론을 이용하여 신뢰도를 분석하고자 한다. 일반화가능도이론을 이용하여 신뢰도를 분석함으로써 두 가지 목적을 달성하고자 하는데, 첫째는, 현재의 신뢰도 측정뿐 아니라 측정조건의 변화에 따른 신뢰도 예측이 가능하므로 현재의 신뢰도가 적정한 수준인지 그렇지 않다면 과제수를 얼마만큼 증가시킬 때 적정 신뢰도가 확보될 수 있을 것인지를 알고자 한다. 두 번째로는, 통상적으로 개인과 문항만을 고려하여 신뢰도를 분석하는데서 나아가 개인이 속한 학교라는 집단도 개인의 점수에 영향을 미치는 하나의 요인으로 간주하고 학생, 학교, 과제의 효과들을 평가하고자 한다. 학교의 효과 크기를 알아냄으로써 개인이 속한 집단을 고려했을 때와 그렇지 않았을 때 신뢰도에 어떤 차이가 있는지 비교할 것이다.

본 연구에서 개인이 속한 그룹, 즉 학교의 효과를 살펴보고자 하는 이유는 다음과 같다. 첫째, 컨소시엄 결과자료를 분석한 선행연구들에서 과제별 분석을 하거나 할 때 학교구분 없이 사례수를 합하여 사용하곤 하지만, 엄밀히 말하면 이는 학교별로 동등하게 무선표집을 한 것이 아니다. 둘째, 컨소시엄에서 실기평가방법은 공유한다 하더라도 각 학교의 임상실기 교육의 내용이나 방법은 여전히 차이가 있을 것이며, CPX 시험의 시행은 일률적으로 같은 시기, 같은 장소에서 시행하는 것이 아니라 각 학교 실정에 따라 과제와 대상을 선별하여 자율적으로 시험을 시행한다. 따라서 과제, 채점자, 측정시기, 측정횟수 등과 마찬가지로 개인이 속한 학교도 개인의 점수에 영향을 주는 하나의 오차원으로 분석하여 그 효과를 알아볼 필요가 있다. 마지막으로 과제와 학교의 상호작용 효과를 알 수 있기 때문이다. 의과대학들이 공유하는 CPX 과제가 특정학교에 유리하거나 불리한 결과를 유도하지 않았는지 그 영향력을 가늠해볼 수 있을 것으로 기대한다.

이와 같은 목적에서 본 연구의 연구문제는 다음과 같다.

첫째, CPX의 오차 요인으로서 개인, 학교, 과제를 설정할 때, 각각의 효과 크기는 어떠한가?

둘째, 현재의 측정 상황에서 CPX의 신뢰도, 즉 일반화가능도계수 및 의존도계수의 크기는 어떠한가?

셋째, 측정 조건이 변화함에 따라 신뢰도는 어떻게 변화하겠는가?

본 연구는 Cronbach's α값에 의존한 단편적인 신뢰도 분

석에서 벗어나 일반화가능도이론을 적용하여 다양한 오차원을 규명하는 신뢰도 분석을 시도한다는 점에 의의가 있을 것이다.

대상 및 방법

1. 연구 자료

연구 자료는 2008년 서울·경기지역 CPX 컨소시엄에 속한 16개 의과대학 학생들을 대상으로 한 임상수행능력평가 결과이다. 이 의과대학들은 컨소시엄 내에서 개발된 문항과 표준화 환자를 활용하여 각 학교의 사정에 맞게 과제와 학생집단(4학년 또는 3학년 등)을 선별하여 시험을 시행하였다. 본 연구 자료에는 모든 학교들이 공통으로 사용한 6개 과제와 동질적인 학생집단(4학년)에 대하여 치른 결과를 대상으로 하였다. 6개 과제는 각각 '1. 피곤해요, 2. 속이 쓰려요, 3. 생리가 많아요, 4. 여기저기가 아파요, 5. 가슴이 두근거려요, 6. 목이 답답해요'라는 주제로 환자를 대면하는 내용이다.

연구 대상이 된 16개 학교, 총 학생 수는 1,319명이다. 각 과제별 점수는 100점 만점으로 환산된 점수이며 각 학교별 점수 평균, 과제별 점수 평균은 Table 1과 같다.

2. 연구 모형 및 분석 방법

일반화가능도이론은 G연구와 D연구로 구분된다. G연구에서는 측정상황에서 발생할 수 있는 다중오차요인을 동시에 분석하고, 측정점수에 대한 오차요인의 상대적 영향력을 산출할 수 있으며, G연구결과를 기초로 D연구에서는 일반화가능도 계수와 함께 의사결정자에게 안정적인 점수를 얻기 위한 측정조건을 제시한다[4]. 일반화가능도이론을 이해하기 위한 개념적 설명 및 모형 설계 원리, 신뢰도 산출방법 등은 관련문헌을 참고하여야 할 것이다[4,5].

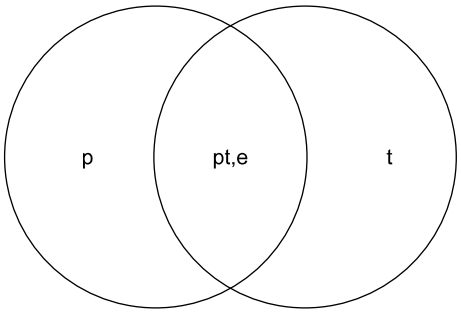
본 연구 자료에서 기본적으로 개인을 측정대상으로, 과제를 국면으로 하는 가장 기본적인 단일국면 교차설계 모형을 가정할 수 있다. 피험자 개인, 즉 학생을 p (student)로, 과제를 t (task)로 표기할 때, p × t로 표현할 수 있으며, 이때 일반화가능도계수는 고전검사이론의 신뢰도 계수, Cronbach's α값과 같다[5].

Table 1. Descriptive Statistics of Data Used in This Study

Univ.	No. of students	Task1		Task2		Task3		Task4		Task5		Task6	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	42	52.7	6.6	58.6	9.1	58.3	6.7	62.4	12.2	59.1	8.0	58.8	9.3
2	118	52.0	8.3	57.6	8.9	54.0	8.1	58.7	11.4	52.6	11.6	62.1	11.2
3	53	45.2	8.2	52.7	8.8	49.3	7.9	56.9	12.6	51.6	10.8	62.5	12.0
4	182	53.8	8.0	59.0	9.3	56.7	8.7	61.3	11.4	55.9	9.9	63.5	10.9
5	30	51.1	9.7	58.7	11.5	55.5	6.9	67.0	11.0	55.1	10.6	61.6	12.4
6	103	50.5	9.6	58.5	9.5	58.9	8.0	59.1	12.3	55.3	10.3	59.3	11.2
7	49	59.9	8.8	66.1	8.5	66.1	7.7	75.8	10.0	61.2	12.2	70.9	8.7
8	133	47.5	7.0	50.8	7.9	52.2	7.6	60.7	10.4	51.4	9.9	59.6	11.6
9	36	53.0	8.3	56.1	10.1	56.3	7.8	58.8	12.4	55.6	7.3	54.0	9.1
10	86	47.5	8.7	54.3	8.9	53.4	6.3	61.5	10.3	54.9	10.6	63.1	11.0
11	101	50.7	7.0	60.8	8.4	54.8	7.9	60.5	10.2	56.2	8.6	60.2	11.8
12	57	45.5	8.0	51.1	9.4	52.0	8.8	53.2	12.0	43.6	10.5	55.2	11.7
13	91	47.2	7.7	54.9	8.8	51.9	7.7	58.9	11.3	47.5	9.5	58.9	12.0
14	67	45.6	7.8	56.7	7.4	55.7	9.4	61.0	11.2	60.1	10.9	62.0	11.6
15	122	49.9	9.0	54.6	9.8	54.2	9.7	62.1	11.4	57.0	10.3	65.2	9.7
16	49	47.9	7.7	56.3	9.8	55.4	7.9	65.1	10.8	55.4	8.7	61.8	9.8
Total	1,319	50.0	8.2	56.7	9.1	55.3	7.9	61.4	11.3	54.5	10.0	61.2	10.9

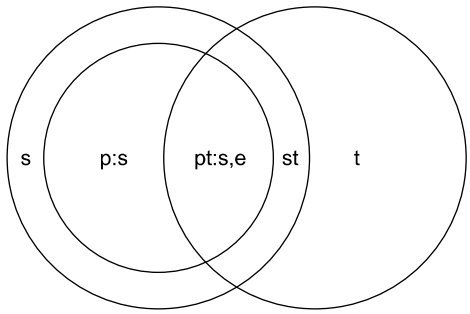
SD: Standard deviation.

Fig. 1. p × t Design



p: Student, t: Task.

Fig. 2. (p:s) × t Design



p: Student, t: Task, s: School.

이 기본적인 모형과 비교하여, 연구자는 (p : s) × t를 연구모형으로 자료를 분석하였다. 이것은 개인이 집단(학교, s)에 내재된 상태를 반영한 것으로서, 이해를 돕기 위해 단일국면 모형과 본 연구모형의 분산성분 구조를 벤다이어그램으로 비교하면 Figs. 1, 2와 같다. Fig. 1은 개인과 문항만을 교차시킨 단일국면 설계이다. 이 때, 학생을 학교에 내재된 개인으로 본다면 학교가 하나의 국면으로 추가되어 Fig. 2와 같은 모형 구조가 된다. 같은 학교 학생은 학교의 교육과정, 교육환경 등에 공통적으로 영향을 받고, 타 학교 학생과는 차별화되는 특성을 가질 수 있으므로 그러한 학교 효과를 분석하고자 한다면 적절한 모형은 (p : s) × t가 된다. 이 모형은 학생이 학교에 내재된 자료구조를 반영한 모형으로서, 기본모형에서 학교라는 국면을 더하여 오차원을 세분화한다. 따라서 개인의 분산 이외에도 학교분산, 과제분산, 학교와 과제의 상호작용 분산의 크기를 알 수 있다.

Fig. 1에서 발생하는 개인의 관찰점수 분산은 $\sigma^2_X = \sigma^2_p + \sigma^2_t + \sigma^2_{pt}$로 개인, 과제, 잔차가 분산의 원천으로 단순하게 분산성분이 구성되지만 Fig. 2에 따라, 연구 모형에서 개인의 관찰점수 분산은 아래와 같은 분산성분의 합으로 구성된다.

$\sigma^2_X = \sigma^2_s + \sigma^2_{p:s} + \sigma^2_t + \sigma^2_{st} + \sigma^2_{pt:s}$

σ^2_s : 학교 분산

$\sigma^2_{p:s}$: 학교에 내재된 개인의 분산

σ^2_t : 과제 분산

σ^2_{st} : 학교와 과제의 상호작용 분산

$\sigma^2_{pt:s}$: 잔차 분산

단일국면에서 개인의 분산은 연구모형에서 개인과 학교의 분산으로 분할되며, 과제 효과도 순수한 과제의 효과와 학교와 과제의 상호작용 효과로 나누어짐을 알 수 있다.

일반화가능도계수의 정의에 따라, 측정대상 p가 그룹에 내재되어 있을 때 일반화가능도계수는 아래 공식과 같이 산출된다. 이때 무선적으로 추출된 그룹 속에 내재한 개인을 측정대상으로 한 일반화가능도계수(1)와, 그룹을 고려하지 않은 개인을 측정대상으로 한 일반화가능도계수(2)는 구별된다[14]. 이것은 일반화가능도계수를 해석하는 데 차이를 가져오며 그룹의 분산에 따라 두 값의 차이가 커질 수도, 작아질 수도 있음을 의미한다. 그룹에 내재한 개인, 즉 학교 내 학생의 오차분산 및 일반화가능도계수로서 해석하고자 한다면 그룹에 내재한 개인의 전집점수공식(2)의 일반화가능도계수는 학교 분산을 개인의 분산에 포함시켜서 결과적으로 단일국면 교차설계의 결과와 같다.

$$E^2_\rho(p:s) = \frac{\sigma^2(p:s)}{\sigma^2(p:s) + \sigma^2(pT:s)} \quad \text{equation (1)}$$

$$E^2_\rho(p) = \frac{\sigma^2(s) + \sigma^2(p:s)}{\sigma^2(s) + \sigma^2(p:s) + \sigma^2(sT) + \sigma^2(pT:s)} \quad \text{equation (2)}$$

자료 분석은 urGENOVA 프로그램으로 수행하였다. urGENOVA는 비균형 무선 효과 설계(unbalanced random effects design)의 분석에 적용된다[15,16]. 요인설계에서 각 항의 사례수가 동일하거나 일정한 비율을 유지하고 있지 않은 경우를 비균형 설계라고 한다[17]. 본 연구 자료에서 각 학교에 내재한 피험자의 사례수가 최소 30명에서 최대 182명까지로 동등하지 않고 그 차이가 크므로 비균형 설계를 적용하여 분석하였다. G연구 결과의 분산 성분을 통해 D연구를 수

행하였으며 과제수를 변화할 때 그에 따른 신뢰도계수의 변화를 추정하였다.

결과

위와 같은 모형으로 G연구를 수행한 결과는 Table 2와 같다. 학교 분산이 차지하는 비율은 전체 분산의 6.2%, 학교에 내재된 개인의 분산은 14.9%, 과제 분산이 14.4%, 학교와 과제의 상호작용이 차지하는 분산이 3%, 잔차가 61.5%를 차지하였다.

만약 학교를 고려하지 않고 단일국면으로 본다면 개인분산이 26.5로 전체의 21%를 차지하게 된다. 그러나 그룹의 효과를 분석함으로써 개인분산 중 학교분산이 7.846을 차지하고 학교에 내재된 개인의 분산이 18.676으로 나누어졌다.

G연구의 결과를 바탕으로 측정조건을 변화에 따른 신뢰도 변화를 예측하였다. 현재 과제수 6개로부터 과제수를 8개, 10개, … 18개까지 증가시켰을 때의 일반화가능도계수와 의존도계수는 Table 3과 같다. Table 3의 좌측은 (p : s) × t의 모형의 D연구 결과이며, 앞에서 제시한 일반화가능도계수를 산출하는 공식에 따라 그룹에 내재된 개인을 고려하지 않은 경우를 오른쪽과 같이 제시하고 그 결과를 비교해 보았다.

Table 3의 두 모형을 비교하면, 집단에 내재한 개인을 가정하지 않으면 현재 6개 과제를 대상으로 했을 때 일반화가능도계수가 0.66이며 0.80 이상의 일반화계수를 얻기 위해서는 13개 이상의 과제가 필요하다. 그러나 집단에 내재한 개인을 대상으로 본다면 현재 6개 과제의 일반화가능도계수는 0.59로 더 낮았으며, 0.80 이상의 일반화가능도계수를 얻기 위해서는 17개까지 과제수를 증가시켜야 할 것으로 나타났다.

Table 3의 결과를 도식화하여 과제수 변화에 따른 일반화가능도계수와 의존도계수의 변화율을 그래프로 나타내면 각각 Figs. 3, 4와 같다.

Figs. 3, 4에서 그룹에 내재한 개인(p : s)을 가정했을 때보다 그룹을 고려하지 않은 개인(p)을 가정했을 때, 일반화계수

Table 2. G-Study Result of (p:s) × t

Effect	Degree of freedom	Variance component	(%)
s	15	7.84606	(6.2)
p:s	1,303	18.67588	(14.9)
t	5	18.09019	(14.4)
st	75	3.83178	(3.0)
pt:s	6,515	77.25912	(61.5)
Total	7,913	125.70303	(100.0)

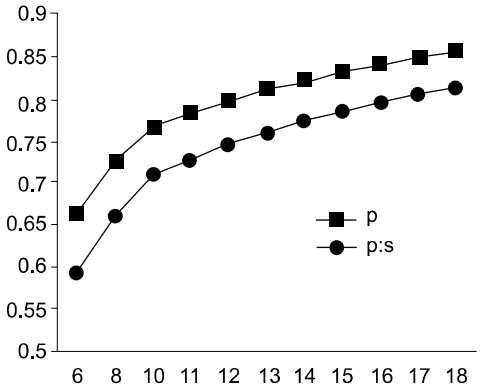
p: Student, t: Task, s: School.

Table 3. Comparison of D-study Result between (p:s) × t Model and p × t Model

Nt	(p:s) × t model					p × t model				
	Universe score	Relative error	Absolute error	g-coefficient	d-coefficient	Universe score	Relative error	Absolute error	g-coefficient	d-coefficient
6	18.67588	12.87652	16.53018	0.592	0.530	26.52194	13.51515	16.53018	0.662	0.616
8	18.67588	9.65739	12.39764	0.659	0.601	26.52194	10.13636	12.39764	0.723	0.681
10	18.67588	7.72591	9.91811	0.707	0.653	26.52194	8.10909	9.91811	0.766	0.728
11	18.67588	7.02356	9.01646	0.727	0.674	26.52194	7.37190	9.01646	0.783	0.746
12	18.67588	6.43826	8.26509	0.744	0.693	26.52194	6.75758	8.26509	0.797	0.762
13	18.67588	5.94301	7.62931	0.759	0.710	26.52194	6.23776	7.62931	0.810	0.777
14	18.67588	5.51851	7.08436	0.772	0.725	26.52194	5.79221	7.08436	0.821	0.789
15	18.67588	5.15061	6.61207	0.784	0.739	26.52194	5.40606	6.61207	0.831	0.800
16	18.67588	4.82870	6.19882	0.795	0.751	26.52194	5.06818	6.19882	0.840	0.811
17	18.67588	4.54465	5.83418	0.804	0.762	26.52194	4.77005	5.83418	0.848	0.820
18	18.67588	4.29217	5.51006	0.813	0.772	26.52194	4.50505	5.51006	0.855	0.828

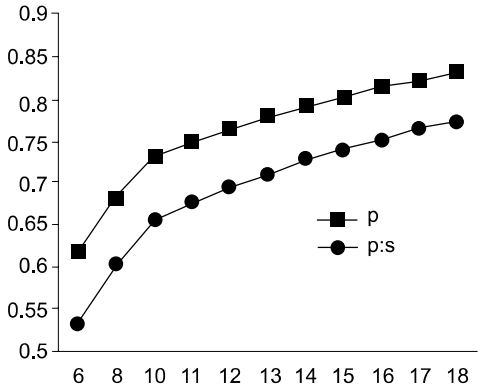
p: Student, t: Task, s: School, g-coefficient: Generalizability coefficient, d-coefficient: Dependability coefficient.

Fig. 3. Generalizability Coefficient of p × t Model vs. (p:s) × t Model



x axis: Number of tasks, y axis: Generalizability coefficient.

Fig. 4. Dependability Coefficient of p × t Model vs. (p:s) × t Model



x axis: Number of tasks, y axis: Dependability coefficient.

와 의존도계수가 일관되게 더 높게 추정되었으며 일반화가능도계수에서보다 의존도계수에서 그 격차가 더 크게 나타났다. 일반화가능도계수는 절대평가에서, 의존도계수는 상대평가에서의 참조할 신뢰도계수이다. 또한 그래프의 증가하는 모양을 보면, 일반화가능도계수와 의존도계수 모두 현재 6개의 과제에서 10개까지 과제수를 증가할 때 계수의 증가율이 높았으며 10개 이상에서는 비교적 완만하게 증가하며 과제수가 증가할수록 두 그래프의 격차는 작아지고 있다.

고찰

본 연구는 일반화가능도이론을 적용하여 CPX 결과의 신뢰도를 분석하였다. 개인의 점수에 영향을 미치는 개인, 학교, 과제 효과를 세분화하여 알아보았으며 결정연구를 통해 과제수가 변화함에 따른 신뢰도 변화를 예측하였다.

Figs. 1, 2의 설계모형의 원리에서, 학교를 고려하지 않고 단일국면으로 본다면 Fig. 1에서의 개인 분산(p)은 Fig. 2에서 학교분산(s)과 학교에 내재한 개인의 분산(p : s)을 합한 것과 같아지므로, 그룹의 효과를 분석함으로써 개인분산 중 학교분산이 차지하는 양을 밝힐 수 있었다. 연구결과 Table 2에서 개인분산 중 학교분산이 6.2%를 차지하고 학교에 내재된 개인의 분산이 14.9%였으나, 이를 단일국면으로 분석하면 개인의 분산이 21%가 되므로, 일반화가능도계수의 산출 원리에 따라 단일국면의 신뢰도가 더 높아진다. 따라서, 이렇게 내재된 자료의 구조를 반영하지 않는다면 개인의 분산이 커져서 신뢰도를 과대추정하게 된다.

각 오차원의 분산성분 비율을 비교하여 볼 때, 잔차를 제외하고 분산의 크기가 개인>과제>학교>학교와 과제의 상호작용 순서로 나타났으며, 특히 학교와 과제의 상호작용이 차지하는 비율이 매우 낮게 나타났다. 이는 개인의 점수에 미치는 학교의 영향력이 과제의 영향력보다 작고, 학교와 과제의 상호작용의 영향력이 매우 낮음을 의미하므로 컨소시엄 내에서 학교의 영향이 개인의 점수분산에는 일정 비율을 차지하지만, 과제와 상호작용 하여서는 큰 영향을 주지 않은 것으로 나타났다. 즉, 학교에 따라서 과제가 다르게 기능하지 않았음을 보여주는 것으로서 컨소시엄에서 개발된 과제들이 특정 학교에 유리함이나 불리함이 없이 고르게 작용했다고 해석할 수 있겠다. 이 결과는 컨소시엄 구성 효과를 가시적으로 입증한 하나의 결과라고 볼 수 있다. 즉, 학생 점수의 차이는 학생 고유의 차이에 기인한 효과가 가장 크고 과제나 학교에 의한 영향이 상대적으로 작으며, 특히 공통으로 적용한 과제들은 컨소시엄 내 모든 학생들에게 공평한 평가도구로서 적절했다고 볼 수 있다. 과제의 분산이 선행연구들보다 작아진 것도 바람직한 현상이다. 단일국면으로 분석했을 때, 개인의 분산 수준이 유사한 가운데 과제의 분산은 선행연구들에서 18%, 21%의 분산을 보였으나[12,13], 본 연구에서는 14%로 작아져 이 또한 컨소시엄이 회를 거듭할수록 공통문항들이 수준에 따른 편차가 감소하는 것으로 보인다.

연구결과 Table 3에 따르면, Brennan [14]이 언급한대로

그룹에 내재한 개인을 가정하지 않고 개인을 측정대상으로 했을 때 신뢰도는 더 높게 추정되었으며, 결과적으로 단일국면 p × t의 D연구 결과와 같다. 따라서 집단에 내재한 개인을 가정하지 않으면 현재 6개 과제를 대상으로 했을 때 일반화가능도계수가 0.66이며 0.80 이상의 일반화계수를 얻기 위해서는 13개 이상의 과제가 필요하다. 이러한 결과는 일반화가능도이론을 적용해서 이전 년도의 CPX 결과를 분석했던 선행연구들과 유사한 결과로 나타났으며[11,12,13], 선행연구들도 개인과 과제만을 고려한 분석 결과였다. 반면, 그룹에 내제된 개인을 고려한다면 6개 과제의 일반화가능도계수는 0.59로 낮아졌으며, 0.80 이상의 일반화가능도계수를 얻기 위해서는 17개 이상의 과제가 필요한 것으로 나타났다.

Figs. 3, 4에서 그룹의 효과를 고려하지 않으면 전 범위에서 일반화계수와 의존도계수가 일관되게 더 높게 추정되었으며 일반화가능도계수에서보다 의존도계수에서 그 격차가 더 컸다. 두 계수 모두 현재 6개의 과제에서 10개까지 과제수를 증가할 때 증가율이 높았으며 10개 이상에서는 비교적 완만하게 증가하며 과제수가 증가할수록 두 그래프의 격차는 작아졌다. 따라서 과제수가 충분히 많다면 그룹의 효과를 고려하는지 여부에 따라 신뢰도 추정의 차이가 미미할 것이지만 그렇지 않다면 만족할 만한 신뢰도 충족을 위한 과제수의 의사결정도 달라질 수 있음을 보여준다. 과제수가 작을수록, 그룹의 효과크기가 클수록 위와 같은 격차는 더 커질 것이므로 일반화하고자하는 대상과 그룹의 효과 크기를 고려하여 그 결과를 해석함이 바람직하다. 학교 효과의 크기가 충분히 작다고 인정되기 전에 컨소시엄 전체의 결과를 일개 학교의 학생들을 대상으로 일반화해서는 안 될 것이다.

컨소시엄 내 학교분산과 과제분산, 학교와 과제의 상호작용 효과가 낮게 나타나 컨소시엄을 통해 개발한 과제가 대상 학교들에게 공정한 평가도구가 되었다는 긍정적인 효과로 평가하였다. 다만, 만족할만한 신뢰도 확보를 위해서는 더 많은 공통과제의 적용이 필요할 것으로 보인다. 현실적으로 과제수를 늘리는 것은 쉽지 않다면 다른 오차원을 줄여 신뢰도를 높일 수 있는 방향이 모색되어야 할 것이다.

기존의 CPX 점수 분석은 컨소시엄 내 전체 피험자를 대상으로 과제의 신뢰도를 계산하거나, 개인의 점수 분석과는 별개로 학교의 평균을 비교하여 그 격차를 알아보는 방식, 또는 일개 학교의 결과만을 대상으로 분석하는 방식 등이 주를 이루었다[18]. 본 연구에서는 개인, 학교, 과제를 동시에 오차원으로 분석함으로써 개인 점수의 오차분산에 영향을 미치는 학교, 과제, 학교와 과제의 상호작용의 영향력을 각각 분석할 수 있었다.

개인의 성취도를 평가함에 있어 개인이 소속한 집단의 영향력을 간과할 수 없다. 모든 의과대학이 표준화된 교육과정을 따른다 하더라고 독특한 학교문화와 교육과정 운영, 교수의 질, 실습량, 시설투자, 정보공유 등에 의한 차이로 학생의 성취에 학교가 미치는 영향력은 그 자체로 연구대상이거나 또는 통제하여야 할 요인으로 여겨지곤 한다[1]. 특히 학교 간 사례수의 차이가 큰 비균형설계의 경우에는 단순한 평균비교나 t통계로 그 차이를 단언할 수 없다. 따라서 자료의 구조를 모형에 그대로 반영시켜 분석하는 것이 정확한 결과를 유도하고 섣부른 일반화를 범하지 않는 한 방법일 것이다. 사회과학은 사람, 조직, 프로그램 등을 대상으로 하기 때문에 엄격하게 측정조건을 통제하는 실험연구가 불가능하다. 따라서 사회과학연구는 종속변수에 관련이 있다고 판단되는 다양한 관련변수들을 통계모형에 포함하여 연구자의 관심이 있는 변수 간 관계에 간섭을 일으키는 변수를 통제하고 변수 간 관계를 밝힐 필요가 있다[19].

본 CPX 시험의 결과는 하나의 컨소시엄 내에서 시행된 결과이지만 학교별 시험 일정은 개별 학교별로 다르게 운영되었다. 다수의 문항 스테이션을 시행하면서 여러 일자에 걸쳐 시험을 시행하였으므로 시험장 복제, 시험 시기 등 시험 환경이 다름으로써 성취도에 영향을 줄 수 있는 가능성을 내포한다. 시험 시기에 따른 정보 유출이나 학생들의 정보 공유가 학교마다 동일한 정도로 통제되었다고 가정하기 어렵다. 따라서 학교 효과를 배재하고 피험자와 과제만을 고려한 신뢰도 분석은 여러 매개 효과를 간과하는 오류를 범할 수 있다. 자료구조를 반영하는 모형의 설정에 따라 신뢰도는 과대 혹은 과소 추정될 수 있으므로 자료구조를 더 명확하게 반영하고 오차원을 세분화한 신뢰도 분석이 필요하다. 본 연구모형에서처럼 개인을 포함하는 그룹이 존재한다면 그것을 반영하여 자료를 분석하는 것이 신뢰도 산출을 정확하게 하고 다양한 오차원을 확인할 수 있게 한다. 결과적으로 연구자에게 주는 정보가 더 많다. 개별학교 단위에서 신뢰도 해석과 컨소시엄

전체 피험자를 대상으로 한 신뢰도를 구별하여 해석할 수 있기 때문이다.

CPX 자료에 일반화가능도이론을 적용하면서 오차원을 더 다양화할 수 없어서 아쉬웠으며 이것이 본 연구의 제한점이기도 하다. 과제에는 연기자 효과와 채점자 효과가 내재한다. 환자 역할을 하는 연기자는 하나의 과제마다 3~5명이 분담하며 이들이 학생들을 할당하여 대면하고 평가하였다. CPX의 특성상 과제는 연기자 및 채점자 효과를 포함할 수밖에 없는데, 연기자를 하나의 국면으로 분리하고자 하였으나 피험자와 교차하지 않는 자료구조 때문에 연기자 또는 채점자 국면을 분석에 포함하지 못하였다. 이러한 제한점은 자연히 후속 연구의 제안으로 이어진다. 연기자, 채점자 국면 등 다양한 측정조건의 효과를 파악할 수 있는 연구 설계를 통해 다양한 국면을 평가하는 시도가 활성화되어야 할 것이다.

CONFLICT OF INTEREST

None.

ACKNOWLEDGEMENTS

Data was provided from Seoul-Gyeonggi CPX Consortium.

REFERENCES

1. Seong TJ. Educational evaluation. Seoul, Korea: Hakjisa; 2002.
2. Seoul-Gyeonggi CPX Consortium [Internet]. CPX consortium; c2004 [updated 2010 February 1; cited 2010 March 5] Available from: http://www.cpx.or.kr.
3. Kim JH. Role and developmental direction of Seoul-Gyeonggi CPX Consortium. Paper presented at: 4th CPX Symposium; 2007 December 7; Hanyang University, Seoul, Korea.
4. Kim SS, Kim YB. Generalizability theory. Seoul: Kyoyookbook; 2001.
5. Brennan RL. Generalizability theory: statistics for social science and public policy. New York, USA: Springer; 2001.
6. Cronbach LJ, Glesser GC, Nanda H, Rajaratnam N. The generalizability of behavioral measurement: the theory of generalizability for scores and profiles. New York, USA: John Wiley; 1972.
7. Nam MH. An application of the generalizability theory to performance assessment. J Educ Eval 1996; 9: 73-93.
8. Kang AN, Lee GM. A generalizability theory approach to investigating the generalizability of performance assessment using student peer reviews. J Educ Eval 2006; 19: 107-121.
9. Kim CH, Lee GM. A generalizability theory approach to investigating rater effects with a national exam composed of constructed-response items. J Curr Eval 2008; 11: 141-159.
10. Roh HR, Kim JK, Hwang JY, Park SB, Lee SW. Experience of implementation of objective structured oral examination for ethical competence assessment. Korean J Med Educ 2009; 21: 23-33.
11. Im H, Kim SS. A study of investigating error sources and reliability for clinical performance examination (CPX). J Educ Eval 2005; 18: 27-46.
12. Lee YM. Investigation of error factor of CPX using generalizability theory. Paper presented at: 2nd CPX Symposium; 2005 December 5; Seoul National University Hospital, Seoul, Korea.
13. Park JH. A psychometric evaluation of CPX in relation to validity and reliability. [dissertation]. [Seoul, Korea]: Ewha Womans University; 2008.
14. Brennan RL. The conventional wisdom about group mean scores. J Educ Meas 1995; 32: 385-396.
15. Brennan RL. Manual for urGENOVA version 2.1. Iowa

City, USA: Iowa Testing Programs, University of Iowa; 2001.
16. Crick JE, Brennan RL. Manual for GENOVA: a gerneralized analysis of variation system. Iowa City, USA: Research and Development Division, American College Testing Program; 1983.
17. Korean Society for Education Evaluation. Educational evaluation thesaurus. Seoul, Korea: Hakjisa; 2004.
18. Kim S, Park SW, Hur Y, Lee SJ. The appropriateness of using standardized patients' (SPs) assessment scores in clinical performance examination (CPX). Korean J Med Educ 2005; 17: 163-172.
19. Kang SJ. Regression analysis. Seoul, Korea: Kyoyookbook; 2003.