문항반응이론과 고전검사이론을 이용한 진료수행시험의 문항 분석

고려대학교 의과대학 의학교육학교실, 의학통계학교실[1], 성공회대학교 소프트웨어공학과[2]

임현선 · 이영미 · 안덕선 · 이준영[1] · 임 형[2]

= Abstract =

Item Analysis of Clinical Performance Examination Using Item Response Theory and Classical Test Theory

Hyun-Sun Lim, MA, **Young-Mee Lee,** MD, MSED, PhD,
Duck-Sun Ahn, MD, MA, FRCSC, **Joon-Young Lee[1],** PhD, **Hyung Im[2],** PhD

Department of Medical Education, Department of Medical Statistics[1], Korea University, College of Medicine
Department of Software Engineering, Sungkonghoe University[2], Seoul, Korea

Purpose: The objectives of this study were: 1) to analyze Clinical Performance Examination (CPX) items using item response theory (IRT) and classical test theory (CTT) and 2) to discuss how to apply and interpret these results in order to improve the quality of CPX items. In addition, we intended to explore statistical procedures in order to merge examination data from several different medical schools.

Methods: The subject of the study was the 2005 CPX examination data from 10 medical schools located in Seoul and the Kyunggi province. For merging data from ten different medical schools, Levene's test for homogeneity of variances was used. Homogeneous group selection was conducted based on ANOVA or Kruskal-Wallis' test and Tukey's multiple comparisons appropriately. The generalized partial credit model was applied to analyze polytomous items and the 2-parameter logistic model was used to analyze dichotomous items.

Results: Data from 8 medical schools were incorporated into the analysis. The result of the discrimination index by IRT was different from that of CTT in both polytomous and dichotomous items. Discrimination index from IRT tended to be lower than that of CTT. Difficulty index of dichotomous items of two models was correlated well with each other. However, for polytomous items, IRT model provided more information than CCT.

Conclusion: We discovered that the CPX items were mostly easy in terms of difficulty index, and the result from IRT and CCT model did not correlated well in the discrimination index. IRT may provide more detailed information for polytomous items, but the checklist and criteria of scoring system should be cautiously reviewed.

Key Words: Clinical performance examination, Item response theory, Classical test theory

교신저자: 이영미, 고려대학교 의과대학 의학교육실, 서울시 성북구 안암동 5가 126-1번지
임 형, 성공회대학 소프트웨어공학과, 서울시 구로구 항동 1-1번지
Tel: 02)920-6098, Fax: 02)928-1647, E-mail: ymleehj@korea.ac.kr

서 론

임상수행능력평가를 위한 서울, 경기 컨소시엄(이하, CPX 컨소시엄)에서는 2004년부터 가입대학 의학과 학생들을 대상으로 표준화환자를 이용한 진료수행시험 (clinical performance examination, 이하 CPX)을 시행하고 있다. CPX 컨소시엄 결성의 주목적은 임상수행능력 평가를 위한 실기 시험의 사례 및 문항 개발 그리고 표준화환자 훈련 등에 소요되는 인적, 재정적 자원을 분담함으로써 안정적이고 효율적인 임상수행평가 시스템을 운영하는 것이다. 또한, 시험 후 결과자료에 대한 공유와 공동연구를 통하여 의학교육에 있어 수행평가의 발전을 도모하는 것도 주요 목적 중에 하나이다.

CPX 컨소시엄의 목적 중에 하나인 시험 자료의 공유와 활용은 개별 대학 단위 자료에서 얻을 수 없는 대단위 표본수의 결과를 이용함으로써 모집단에 좀 더 가까운 결과를 추론할 수 있다는 장점이 있다. 그러나 교육과정과 상황이 상이한 각 대학의 자료를 단순히 통합하거나 비교하는 것에 대해서는 이견이 많다. 따라서 각 대학의 시험 결과 자료를 통합하고 의미 있는 양적 자료로 활용하기 위해서는 과학적이고 합리적인 접근법이 필요하다.

2010년 의사국가시험에 임상진료시험의 도입이 확실시 되면서 각 대학 별 혹은 컨소시엄 별 진료수행시험 문항 개발 및 시험 운영, 사후 평가가 관심의 대상이 되고 있다. 임상수행평가의 경험이 아직 풍부하지 않은 우리나라의 경우 신규 문항개발도 시급하지만 이미 출제된 문항에 대한 분석을 시행하여 출제자들에게 피드백을 주고 지속적으로 문항의 질적 향상에 노력하는 것도 동시에 이루어져야 한다. 시험을 구성하는 문항의 양호도에 따라 시험의 신뢰도와 타당도는 영향을 받게 된다 (Hwang, 1999). 따라서 어떤 종류의 시험이든 시험을 구성하고 있는 문항을 분석하여 그 결과를 토대로 문항의 질을 판단하고 문제점이 도출된 문항에 대한 원인 교정을 통하여 해당 시험의 질을 향상 시킬 수 있다. 특히 졸업시험이나 자격시험과 같은 고부담 검사 (high stake examination)의 경우 높은 타당도, 신뢰도, 공정성을 요구하기 때문에 다양한 심리측정학적 분석이 요구되며 사후 문항분석도 이러한 목적으로 활용될 수 있다 (Lee, 1992).

문항분석을 위해서는 고전문항검사이론과 문항반응이론을 사용할 수 있다. 고전검사를 이용한 문항분석은 이해하기 쉽고 적용하기 쉽다는 장점을 지니고 있어 현재 대부분 시험 분석에 활용되고 있다. 그러나 고전검사이론은 집단 특성에 따라 문항 난이도와 변별도가 달라지는 것과 시험 특성에 따라 피험자의 능력을 다르게 추정한다는 단점을 지니고 있다. 이에 비해 문항반응이론은 검사를 치른 집단에 무관하게 문항의 난이도와 변별도를 산출할 수 있으며 학생 자신의 고유한 능력점수를 추정할 수 있다는 장점을 지니고 있다.

현재 CPX 컨소시엄에서 채택하고 있는 채점 방식 중에는 0점과 1점으로 구분되는 이분문항이외도 피험자의 수행능력의 완전 및 불완전성에 따라 2점 (제대로 했음), 1점 (제대로 못했음), 0점 (하지 않았음)의 부분점수를 부여하는 다분 문항이 있다. 이는 임상실기시험의 특성상 내용타당성 측면에서 반드시 필요하다. 그러나 문항에 따라서는 '제대로 했음'과 '제대로 못했음'을 나누는 기준이 주관적이거나 인위적인 경우가 있다. 또한, 이분문항에 비하여 채점기준은 객관적이고 타당하게 설정되어 있다고 하더라도 채점자가 이를 제대로 해석하지 못하거나 일관성 있게 채점하지 못하는 오류가 발생할 소지가 있다. 따라서 부분점수를 부여하는 다분문항에 대한 내용타당성과 신뢰도와 더불어 좀 더 다양한 정보가 제공된다면 문항을 검토하고 개선하는데 도움이 될 것이다. Park (2001)는 다분문항분석이론을 이용할 경우 부분점수 채택 문항의 특성을 좀 더 구체적으로 파악하고 채점기준표를 점검하거나 채점과정을 점검할 필요가 있는 문항을 파악할 수 있는 장점이 있다고 주장하였다.

이 연구의 목적은 2005년도 CPX 컨소시엄의 시험문제를 문항반응이론과 고전검사이론을 이용하여 비교 분석하고 그 결과를 문항 점검 및 향상에 어떻게 활용할 수 있을지를 논의하는 것이다. 아울러, CPX 컨소시엄을 통하여 얻을 수 있었던 여러 대학

Table I. Subject: Numbers of Examinee and Stations of Each Medical School

Name of university*	Examinee (No.)	Station (No.)
A	63	5
B	37	6
C	41	6
D	70	6
E	79	6
F	113	6
G	124	6
H	44	8
I	56	8
J	144	10

* Universities arbitrarily designated with alphabets by authors.

의 자료를 과학적으로 통합하는 통계적 방법을 제시하는 것이다.

대상 및 방법

가. 연구 대상

CPX 컨소시엄 소속 18개 의과대학 중 2005년 7월 22일부터 10월 9일에 걸쳐 실제 임상수행능력 평가를 실행했던 13개 의과대학의 시험 결과를 분석 자료로 사용하였다[1]. 이 중 시험의 목적이 학생의 총괄평가가 아니거나 사용한 사례[2]수가 5개 미만인 3개 의과대학을 분석 대상에서 제외하고 총 10개 대학의 자료를 사용하였다 (Table I).

나. 연구 방법

1) 자료통합 및 신뢰도

사례별로 10개 대학의 자료를 통합하기 위해 각 대학 자료의 분산 동질성과 대학 간 평균 시험점수의 차이를 검증하여 통합 여부를 결정하였다. 사례별로 10개 대학 자료의 Levene의 분산 동질성 검정을 시행하여 분산의 동질성이 검증된 사례들의 경우, 각 대학 간 평균 시험점수의 차이는 일원분산분석 (ANOVA) 및 Tukey 사후검증을 실시하였다. 분산의 동질성이 확보되지 않은 사례들의 경우, 비모수적 방법인 Kruskal-Wallis 검정과 자료의 순위에 근거한 Tukey 사후검증을 시행한 결과에 따라 자료를 통합하였다. 통합된 자료의 각 사례의 신뢰도는 사례를 구성하는 문항 간 내적 일관성 신뢰도 추정 방법 중 Cronbach α값으로 구했다. 통계분석은 SPSS 12.0 프로그램을 사용하였으며, 분석 결과의 통계적 유의성은 유의수준 5%하에서 판단하였다.

2) 문항분석 모형

문항반응이론과 고전검사이론을 이용하여 각 사례의 문항에 대한 난이도와 변별도를 계산하였다. 문항반응이론을 적용한 경우 이분문항 (0점과 1점으로 채점하는 문항)은 2모수 로지스틱 모형 (2-parameter logistic model)으로 변별도와 난이도를 추정하였다. 다분 문항 (0, 1, 2점으로 채점하는 문항)은 일반화부분점수모형 (generalized partial credit model)을 사용하여 각 문항 당 1개의 변별도와 2개의 난이도를 추정하였다. 즉, 다분 문항은 각 부분점수에 해당하는 난이도를 각각 산출할 수 있으며, 0점 이상을 받을 난이도, 1점 이상을 받을 난이도로 해석 할 수 있어서 이를 문항범주 난이도라고 부른다. 문항모수를 추정하기 위하여 프로그램 Parscale 4.1을 사용하였다. 고전검사이론을 이용한 문항분석 방법 중에서 변별도 지수는 양분점 상관계수를 활용하였으며, 난이도는 전체 피험자 중 정답을 맞힌 피험자의 비율을 사용하였다.

a. 변별도 분류 기준

문항변별도 지수를 사용하여 문항을 평가하는 절대적 기준은 없지만 2모수 로지스틱모형을 사용한 문항분석에서는 문항 변별도가 0.34 이하인 문항은 변별도가 거의 없는 문항, 0.35~0.64는 변별도가 낮은 문항, 0.65~1.34는 변별도가 적절한 문항, 1.35~1.69는 변별도가 높은 문항, 1.70이상은 변별도가 매우 높은

1) 본 연구는 자료의 사용에 대하여 CPX 컨소시엄 평가소위원회 심의를 득했음.

2) CPX를 구성하는 각 문항을 시험실 혹은 스테이션으로 부르고 각 시험실은 20개~30개 하위문항 (의사-환사 관계 포함)으로 구성되어 있다. 본 연구에서는 시험실을 사례라고 지칭하고 각 사례별 하위문항을 문항으로 기술하였다.

Table II. Levene's Test for Homogeneity of Variances

	Levene statistic	Numerator df	Denominator df	p-value
Case 1	2.165	9	760	.023
Case 2	4.327	9	760	<.0001
Case 3	2.699	9	760	.004
Case 4	1.761	9	759	.072
Case 5	1.326	9	759	.219
Case 6	4.354	8	698	<.0001
Case 7	0.150	2	240	.861
Case 8	0.401	2	240	.670

문항으로 간주된다 (Seong, 2005). 본 연구에서는 편의상 문항의 변별력을 3개의 구간 즉, 문항변별도 지수가 0.34이하이면 변별력이 거의 없는 문항, 0.35~1.69이면 변별력이 어느 정도 있는 문항, 1.70 이상이면 변별력이 매우 높은 문항으로 분류하였다.

고전검사이론에서 문항의 변별도를 평가하는 기준 역시 여러 가지가 있을 수 있지만 일반적으로 많이 사용하는 분류방법에 따라 0.20 미만인 문항은 변별력이 거의 없어서 수정하거나 제거해야 할 문항, 0.21~0.40 미만이면 변별력 있는 문항, 0.40 이상이면 변별력이 높은 문항으로 간주하였다.

b. 난이도 분류 기준

문항반응이론에서 난이도는 0을 기준으로 어렵고 쉬움이 결정된다. 일반적으로 난이도가 −0.5 미만인 문항은 쉬운 문항, −0.5~0.5인 문항은 난이도가 중간정도인 문항으로, 0.5 이상인 문항은 어려운 문항으로 간주된다 (Seong, 2005).

고전검사이론에서 문항 난이도는 정답률을 뜻하며, 난이도가 0.8 이상인 문항은 쉬운 문항, 0.2~0.8 사이의 문항은 중간 정도, 0.2 미만인 문항은 어려운 문항으로 분류한다.

결 과

가. 대학 간 동질성 검증과 자료통합

각 사례별로 10개 대학의 대학 간 자료의 분산 동질성 검증을 시행하였다. 총 10개 사례 중 사례 9와 10은 1개교에서만 시행하였음으로 분석에서 제외하였다. 분석 결과 사례 4, 5, 7, 8은 대학별 자료의 분산이 같았고 사례 1, 2, 3, 6의 대학별 자료의 분산은 서로 다른 것으로 나타났다 (Table II).

분산의 동질성이 확인된 사례 4, 5, 7, 8의 대학 간 평균 시험점수의 차이는 분산분석과 Tukey 사후검증을 시행하였다. 그 결과 사례 4의 경우, 전체 10개 대학 중 7개 대학 간의 평균은 유의한 차이가 없었으나, A대학, E대학, I대학 간의 평균의 차이는 유의하였다. 따라서 사례 4의 경우 A, E, I 대학의 자료를 제외한 나머지 7개 대학의 자료를 통합하여 연구에 사용하였다. 동일한 방법으로 사례 5는 B, E, F 대학 자료를 제외한 나머지 7개 대학의 자료를 그리고 사례 7과 8은 H 대학 자료를 제외한 나머지 9개 대학의 자료를 통합하여 사용하였다 (Table III). 대학 간 분산의 차이를 보였던 사례 1, 2, 3, 6에 대해서는 Kruskal-Wallis 검정 및 자료의 순위에 기초한 Tukey 사후검증을 실시하였다. 사후검증 결과, 유의한 차이를 보인 대학들의 자료를 제거하고 통합하였다. 즉, 사례 1에서는 A, E, F, H 대학 자료가 제외되었으며, 사례 2, 3, 6에서는 F 대학 자료를 제외하였다.

F 대학의 사례별 평균값은 8개 사례 중 6개 사례가 다른 9개 대학의 평균값과 유의한 차이를 보여 타 대학의 시험 자료와는 이질적인 요소가 있는 것으로 추정하고 모든 자료 통합에서 삭제하였다. 각 사례의 통합된 대학 자료의 최종결과는 Table III과 같다.

Table III. Homogeneous Group Selection Based on ANOVA or Kruskal-Wallis' Test and Tukey's Multiple Comparisons Appropriately

	Method used for group comparisons	df	F-value	p-value	Heterogeneous groups*	Number of schools combined
Case 1	Kruskal−Wallis	9	6.967	<.0001	A, E, F, H	6
Case 2	Kruskal−Wallis	9	8.507	.004	F	9
Case 3	Kruskal−Wallis	9	8.412	<.0001	F	9
Case 6	Kruskal−Wallis	8	12.185	<.0001	F	9
Case 4	ANOVA	9	12.787	<.0001	A, E, I	6
Case 5	ANOVA	9	8.859	<.0001	B, E, F	7
Case 7	ANOVA	2	6.614	.002	H	8
Case 8	ANOVA	2	7.165	.001	H	8

* Identified by Tukey's multiple comparison test.

Table IV. Reliability and the Number of Total, Dichotomous and Polytomous Items of Each Case

	Examinee (N)	Cronbach α	Total Item (N)	Dichotomous (N)	Polytomous (N)
Case 1	594	.505	21	17	4
Case 2	730	.615	19	15	4
Case 3	727	.528	21	10	11
Case 6	584	.554	21	16	5
Case 4	571	.517	17	12	5
Case 5	666	.534	21	20	1
Case 7	184	.563	16	16	-
Case 8	187	.413	12	12	-
			148	118	30

나. 시험 각 사례의 신뢰도

사례별로 통합된 자료를 사용하여 Cronbach α값을 산출하였다. 각 사례별 신뢰도 수준은 0.413~0.615이었으며 각 사례의 응시자 수, 문항 수는 Table IV와 같다.

다. 문항분석결과 비교

각 사례별 문항분석은 앞서 기술한 과정을 거쳐 총 10개의 사례 중 일개 대학에서만 사용한 두 개 사례를 제외한 8개의 사례가 사용되었다. 각 사례는 최소 12개부터 21개까지의 하위 문항으로 구성되어 있었다[3]. 8개 사례의 총 문항은 148개이며 이 중 이분문항 (0점과 1점으로 채점)은 118개이고, 다분문항 (0점, 1점, 2점으로 부분점수가 있는 문항)은 30개였다. 이분문항 중 한 개 문항은 정답자가 한 사람도 없어서 분석 대상에서 제외하여 최종적으로 총 147개 문항을 분석하였다. 8개 각 사례별로 변별도를 3개 수준, 난이도를 3개 수준으로 분류하여 정리한 결과는 Table V와 같다.

1) 다분문항분석 (Table VI)

a. 변별도

문항반응이론의 일반화부분점수모형을 사용하여 다분문항에 대한 1개의 변별도와 2개의 난이도범주를 추정하였다. 그 결과 전체 30개의 다분문항 중 17개 (56.6%) 문항의 변별도가 낮거나 매우 낮았으며, 변별도가 매우 높은 문항은 2개 (6.7%)였다 (Table VI).

3) 각 사례 중 공통적으로 들어가는 환자-의사관계 문항은 임상적 지식과 술기를 평가하는 영역과는 다른 성격을 띤 문항으로 판단되어 이번 연구에서는 제외하였다.

Table V. Classification of Each Case According to Discrimination and Difficulty Indices using Item Response Theory and Classical Test Theory

Case (number of items)	Discrimination index			Difficulty index			
		CTT* (%)	IRT† (%)		CTT (%)	IRT1‡ (%)	IRT2§ (%)
1 (N=21)	Need to be corrected or eliminated	2 (9.6)	9 (42.9)	Easy	5 (23.8)	12 (57.1)	4 (100.0)
	Acceptable	15 (71.4)	12 (57.1)	Moderate	15 (71.4)	4 (19.1)	0 (0.0)
	High	4 (19.0)	0 (0.0)	Difficult	1 (4.8)	5 (23.8)	0 (0.0)
2 (N=19)	Need to be corrected or eliminated	2 (10.5)	6 (31.6)	Easy	6 (31.6)	12 (63.2)	2 (50.0)
	Acceptable	13 (68.4)	10 (52.6)	Moderate	12 (63.2)	3 (15.8)	1 (25.0)
	High	5 (21.1)	3 (15.8)	Difficult	1 (5.2)	4 (21.0)	1 (25.0)
3 (N=21)	Need to be corrected or eliminated	2 (9.5)	11 (52.3)	Easy	9 (42.5)	10 (47.6)	11 (100.0)
	Acceptable	14 (68.4)	10 (47.7)	Moderate	11 (52.3)	3 (14.3)	0 (0.0)
	High	4 (21.1)	0 (0.0)	Difficult	1 (5.2)	8 (38.1)	0 (0.0)
4 (N=21)	Need to be corrected or eliminated	3 (14.3)	5 (23.8)	Easy	8 (38.1)	10 (47.6)	0 (0.0)
	Acceptable	15 (71.4)	16 (76.2)	Moderate	7 (33.3)	0 (0.0)	1 (25.0)
	High	3 (14.3)	0 (0.0)	Difficult	6 (28.6)	11 (52.4)	3 (75.0)
5 (N=17)	Need to be corrected or eliminated	1 (5.9)	6 (35.3)	Easy	5 (29.4)	6 (35.0)	0 (0.0)
	Acceptable	11 (64.7)	11 (64.7)	Moderate	10 (58.8)	5 (30.0)	1 (100.0)
	High	5 (29.4)	0 (0.0)	Difficult	2 (11.8)	6 (35.0)	0 (0.0)
6 (N=21)	need to be corrected or eliminated	3 (14.3)	7 (33.4)	Easy	11 (52.5)	16 (76.2)	5 (100.0)
	acceptable	14 (66.7)	12 (57.1)	Moderate	9 (42.6)	1 (4.8)	0 (0.0)
	high	4 (19.0)	2 (9.5)	Difficult	1 (4.9)	4 (19.0)	0 (0.0)
7 (N=15)	Need to be corrected or eliminated	3 (20.0)	0 (0.0)	Easy	6 (40.0)	9 (60.0)	-
	Acceptable	4 (26.7)	13 (86.7)	Moderate	6 (40.0)	0 (0.0)	-
	High	8 (53.3)	2 (13.3)	Difficult	3 (20.0)	6 (40.0)	-
8 (N=12)	Need to be corrected or eliminated	2 (16.6)	2 (16.6)	Easy	6 (50.0)	5 (41.7)	-
	Acceptable	5 (41.7)	10 (83.4)	Moderate	6 (50.0)	4 (33.3)	-
	High	5 (41.7)	0 (0.0)	Difficult	0 (0.0)	3 (25.0)	-

* CTT stands for classical test theory, † IRT stands for item response theory, ‡ IRT^1 means difficulty index results from IRT for dichotomous items, § IRT^2 means difficulty index results from IRT for polytomous items.

고전검사이론으로 분석한 결과 전체 30개 문항 중 변별도가 낮아 문항을 수정하거나 삭제해야 하는 문항은 1개 (3.3%)였고, 수용 가능한 변별도를 지닌 문항이 15개 (50.0%), 변별도가 높은 문항이 14개 (46.7%)인 것으로 나타났다.

문항반응이론과 고전검사이론의 결과를 비교하기 위하여 산포도를 그린 결과는 Fig. 1과 같다. 문항반응이론을 사용하여 추정한 변별도는 고전검사이론으로 계산한 변별도보다 대부분 낮은 경향을 보여주었다.

Table VI. Comparison of Discrimination and Difficulty Indices of Polytomous Item According to Item Response Theory and Classical Test Theory

		Discrimination index		Difficulty index			Frequency*		
		Item response theory	Classical test theory	Item response theory		Classical test theory	0	1	2
				score 1	score 2				
Case 1	Q16†	0.34	0.22	−2.76	−0.55	0.71	58	224	312
	Q17†	0.60	0.21	−4.40	−0.83	0.83	5	192	397
	Q18	0.29	0.25	−0.59	−2.84	0.77	89	98	407
	Q19	0.29	0.11	2.13	−5.38	0.78	112	32	450
Case 2	Q14	0.35	0.52	0.18	−0.21	0.5	260	204	266
	Q15	0.61	0.59	−0.62	−1.55	0.77	114	106	510
	Q16	0.48	0.60	0.22	−0.89	0.58	233	145	352
	Q17‡	0.34	0.27	4.38	1.80	0.08	642	58	30
Case 3	Q10	0.15	0.36	3.56	−3.69	0.51	299	120	308
	Q11‡	0.15	0.29	6.33	−1.77	0.28	479	93	155
	Q12	0.17	0.25	−0.96	−6.67	0.85	70	85	572
	Q13	0.20	0.33	1.03	−6.68	0.83	96	61	570
	Q14	0.40	0.41	0.77	−5.15	0.91	58	22	647
	Q15	0.23	0.40	1.25	−2.00	0.55	253	143	331
	Q16	0.63	0.48	−0.38	−2.05	0.81	101	70	556
	Q17	0.71	0.36	−0.91	−3.19	0.95	27	22	678
	Q18	0.27	0.34	−1.39	−4.09	0.86	59	91	577
	Q19‡	0.20	0.41	7.53	−10.06	0.68	224	17	486
	Q9	0.44	0.42	0.14	−1.98	0.71	162	103	462
Case 4	Q13‡	0.27	0.36	5.64	0.41	0.11	496	42	46
	Q14‡	0.26	0.33	6.93	−0.02	0.09	519	28	37
	Q15	0.57	0.54	1.54	1.13	0.22	402	106	76
	Q16†	0.66	0.52	1.12	1.95	0.19	402	138	44
	Q17†	0.44	0.42	−4.99	0.65	0.69	402	336	237
Case 5	Q16‡	0.19	0.42	6.09	0.28	0.17	441	63	67
Case 6	Q15†	4.27	0.65	−1.82	−0.52	0.84	18	177	471
	Q16†	3.82	0.66	−1.82	−0.53	0.84	19	174	473
	Q17‡	0.12	0.37	4.13	−7.28	0.63	203	88	375
	Q18‡	0.10	0.40	11.73	−5.51	0.29	443	62	161
	Q19‡	0.13	0.28	15.78	−5.05	0.10	589	18	59

* frequency: number of examinees acquired each score, † items which had the difficulty index of score 2 were higher than that of score 1, ‡ items which had the difficulty index of score 1 were greater than 4.

b. 난이도

일반화부분점수모형을 사용하여 다분문항 분석을 하여 각 문항에 대하여 1점을 받을 난이도(범주난이도 1)와 2점을 받을 난이도(범주난이도 2)를 추정하였다 (Table VI).

Table VI를 보면 총 30개의 다분문항 중 24개 문항에서 범주난이도 1이, 범주난이도 2 보다 큰 값을 보였다. 이는 1점을 받는 것이 2점을 받는 것보다

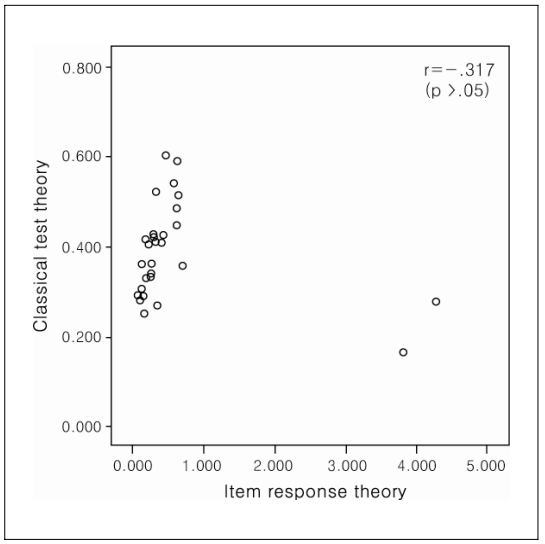

Fig. 1. Scatter plot of discrimination index for polytomous items.

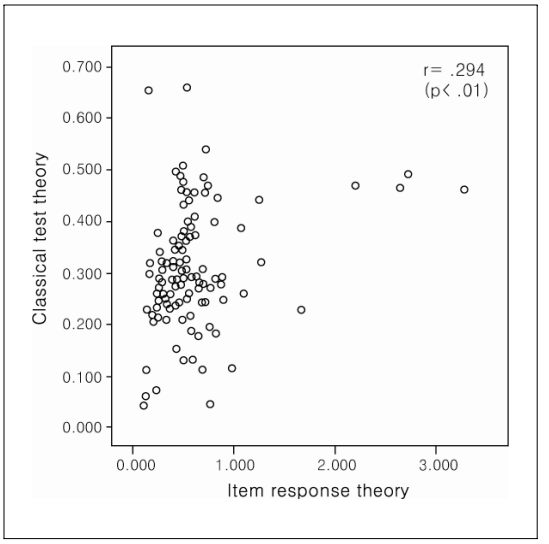

Fig. 2. Scatter plot of discrimination index for dichotomous items.

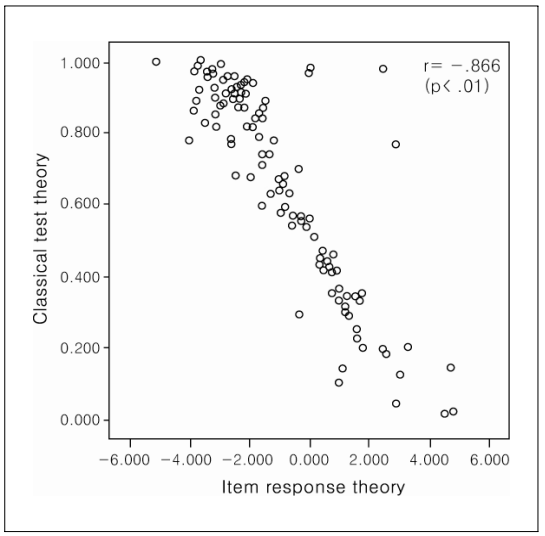

Fig. 3. Scatter plot of difficulty index for dichotomous items

훨씬 어렵다는 것을 의미한다. 이를 사례별로 살펴보면, 사례1의 2개 문항 (Q18, Q19), 사례2의 4개 문항, 사례3의 11개 문항, 사례4의 5문항 중 3문항 (Q13, Q14, Q15), 사례5의 1문항, 사례6의 3개 문항 (Q17, Q18, Q19)에서 범주난이도 1의 값이 범주난이도2의 값보다 컸다. 특히 범주난이도 1의 값이 4.0이상으로 큰 값을 갖는 문항은 9개 문항 (사례2의 Q17, 사례3의 Q11과 Q19, 사례4의 Q13과 Q14, 사례5의 Q16, 사례6의 Q17, Q18, Q19)이었다. 이러한 문항들에 대하여 수험자들이 0점, 1점, 2점을 받은 빈도를 비교한 결과, 0점을 받은 빈도가 1점을 받은 빈도보다 훨씬 높았고, 2점을 받은 빈도도 1점을 받을 빈도보다 높아서 대부분 1점을 빈도가 가장 낮은 것으로 나타났다.

사례1의 Q16과 Q17, 사례4의 Q17, 사례6의 Q15과 Q16처럼 문항이 쉽고, 2점의 빈도, 1점의 빈도, 0점의 빈도가 순서대로 높으면 범주난이도 1이 범주난이도 2보다 낮아서 부분점수가 의미가 있으며, 다분문항의 변별력이 어느 정도 있는 것으로 나타났다. 특히 사례6의 Q15, Q16의 변별도는 4.27과 3.82로 가장 변별력이 높은 문항이었다. 또한 사례4의 Q16은 어려운 문항으로 0점의 빈도, 1점의 빈도, 2점의 빈도가 순서대로 높으며 범주난이도 2가 범주난이도 1보다 높아서 역시 부분점수가 의미가 있고, 변별력이 있는 다분문항인 것으로 나타났다.

고전검사이론으로 계산한 다분문항의 난이도는 각 다분문항 점수의 합을 만점으로 나누어 난이도

를 계산하였다. 사례2의 Q17, 사례4의 Q13, Q14, Q16, 사례5의 Q16번 문항, 사례6의 Q19는 매우 어려운 문항으로써 정답에 반응한 학생이 20% 미만이었다. 이 6개 문항을 제외한 16개 문항의 난이도는 대부분 0.6~0.8 사이에 분포하여 다분문항은 전반적으로 쉬운 것으로 나타났다.

2) 이분문항 분석

a. 변별도

문항반응이론을 적용하여 총 117개 이분문항의 변별도를 구한 결과, 33개 (28.2%)의 문항이 변별도가 거의 없는 문항으로 나타났다. 변별도가 있는 문항은 80개 (68.4%)였고 높은 변별도를 보인 문항은 4개 (3.4%)였다.

이에 반해, 고전검사이론을 적용한 경우, 총 117개 문항 중 16개 (13.7%)문항이 변별도가 거의 없어서 수정하거나 제거해야 하는 문항으로 나타났다. 77개 (65.8%)문항은 변별도가 있는 문항이며, 변별도가 높은 문항은 24개 (20.5%)였다.

문항반응이론과 고전검사이론으로 계산한 변별도를 산포도 (Fig. 2)로 비교한 결과, 두 가지 방법에 의한 변별도의 분포가 상이하였으며, 전체적으로 문항반응이론으로 계산된 변별도가 고전검사에서 산출된 값보다 낮은 경향을 보였다.

b. 난이도

117개의 이분문항에 대하여 문항반응이론을 적용하여 분석한 결과, 쉬운 문항은 47개 (40.2%), 중간수준의 문항은 59개 (50.4%), 어려운 문항은 11개 (9.4%)였다. 고전검사이론을 적용하여 분석한 결과는 쉬운 문항이 70개 (59.8%), 중간수준의 문항이 16개 (13.7%), 어려운 문항이 31개 (26.5%)로 나타났다. 두 가지 방법에 의한 난이도 분석 결과의 산포도는 Fig. 3과 같으며 두 가지 분석방법 상관성은 상관계수가 −.866 ($p<0.01$)로 매우 높은 것으로 나타났다.

고 찰

이 연구는 대규모 자료의 통합 방법과 문항 점검 및 향상을 위한 방법을 모색하고자 하는 목적으로, 2005년에 시행된 서울과 경기도에 위치하는 CPX 컨소시엄 소속 의과대학 중 10개 의과대학의 CPX 시험 자료를 통합한 후 총 8개 사례를 문항반응이론과 고전검사이론을 활용하여 문항분석을 실시하였다.

저자들은 문항분석에 앞서 각 대학의 자료를 과학적이고 합리적인 방법으로 통합하는 방법을 모색해보았다. 서로 다른 집단에서 얻어진 자료를 비교하기 위해서는 점수의 분포가 서로 비슷한 모양을 이룰 때 적절한 비교가 될 수 있다 (Munro, 2001). 이에 분산 동질성과 대학 간 평균 시험점수의 차이를 검정하여 통합 여부를 결정하였다. 10개 대학 자료의 Levene의 분산 동질성 검증을 시행하였고 분산의 동질성이 검증된 사례들은 일원분산분석 (ANOVA) 및 Tukey 사후검증을 실시하였다. 분산의 동질성이 확보되지 않은 사례들은 비모수적 방법인 Kruskal-Wallis 검정과 자료의 순위에 근거한 Tukey 사후검증을 시행한 결과에 따라 자료를 통합하였다. 물론 본 연구에서 사용한 자료 통합 방법 이외도 다양한 방법으로 과학적 자료 통합이 가능하겠지만 연구자들이 사용한 방법이 향후 단위 대학 간 평가 자료를 통합하는 한 가지 방법으로 고려될 수 있을 것으로 생각된다.

문항반응이론과 고전검사이론을 적용하여 문항의 변별도를 분석한 결과는 차이를 나타냈다. 즉, 다분문항 (6개 사례의 총 30개 문항)의 경우, 다분문항반응모형을 적용하면 17개 문항의 변별도가 매우 낮아서 학생들의 능력수준을 제대로 변별하기 어려운 문항으로 나타났으나 고전검사이론을 적용하면 1개 문항만이 변별력이 없는 것으로 나타났다. 이분문항의 경우도 두 가지 검사간의 변별도 결과가 차이를 보였고 문항반응이론으로 산출한 변별도가 고전검사결과보다 낮은 경향을 보였다. 이는 고전검사이론만으로 CPX 문항의 변별도를 계산할 경우 변별력이 과대평가될 가능성이 있음을 추론할 수 있게 한다. 이 결과는 문항반응이론과 고전검사이론의 문항분석의 결과가 상관관계가 높았다고 보고한 기존의 연구 결과 (Lee, 1992; Lee, 1993; Lim *et al.*, 2004)와는 일치하지 않는다. 기존의 연구가 대부분 지필

고사의 이분문항에 대한 분석이었는데 반하여 본 연구의 대상은 수행형 문항이며 또한 다분문항을 포함하고 있다는 것이 이러한 차이를 부분적으로 설명할 수 있게 한다.

난이도의 추정에 있어서는 이분문항의 경우 문항반응이론과 고전검사이론을 이용한 분석 결과는 전체적으로 쉬운 것으로 나타났으며 두 가지 방법 사이의 상관성이 높았다. 다분문항의 경우는 다분문항반응이론을 적용하였을 때 고전검사이론에서는 추정할 수 없는 부분점수의 난이도를 추정할 수 있게 함으로써 좀 더 풍부한 정보를 얻을 수 있었다. 즉, 2점을 받을 난이도와 1점을 받을 난이도 추정치를 비교한 결과 30개 문항 중 24개가 범주난이도 1보다 범주난이도 2가 높은 것으로 나타났다. 이 결과는 다분문항이 아주 어렵거나 쉬운 경우에 자주 나타났다. 쉬운 다분문항에서는 대부분의 학생이 만점인 2점을 받고 1점을 받은 학생은 0점을 받은 학생들보다 훨씬 적었다. 어려운 문항은 0점을 받은 학생이 대부분이며 1점을 받은 학생보다 2점은 받은 학생이 역시 많은 경우가 대부분이었다.

일반적인 시험에서는 능력이 낮은 학생이 0점, 능력이 중간인 학생이 1점 그리고 능력이 높은 학생이 2점을 받아야 한다는 것이 부분점수를 채택하는 시험의 기본적 가정이다. 따라서 대부분 학생이 0점이나 2점을 받고, 1점을 받는 학생은 적어서 중간수준의 능력을 가진 학생들을 제대로 변별할 수 없는 문항이라면 부분점수를 부여하는 것을 제고해야 할 필요가 있다. 즉, 이런 현상이 발생하는 이유는 문항이 지나치게 어려워서 반응을 못하는 학생이 많기 때문이거나, 또는 부분점수 1점을 주는 채점 기준이 적절하지 않기 때문일 수 있어서 채점표나 채점과정을 점검하는 것이 필요하다 (Park, 2001).

CPX는 일반적인 시험과 달리 수험자들의 능력을 변별하지 못한다 해도 목적이나 내용상 반드시 부분점수제를 채택하는 것이 타당할 수 있다. 즉, 신체진찰, 병력청취, 정보 나누기 등의 영역에서 표준적 절차에 따라 완벽하게 시행하는 것과 시늉만 내는 것은 엄격한 차이가 있기에 부분점수의 부여가 중요하다. 따라서 일반적인 수행평가와는 달리 주의 깊게 다분문항반응이론의 결과를 해석해야 할 것이다.

흥미로운 사실은 이 시험에서 사용되었던 채점기준표를 연구자들이 내용검토를 한 결과, 다분문항이론 분석에서 부분점수부여 하는 것이 중간능력수준의 학생을 변별하는 데 크게 기여하지 못하는 것으로 분석된 문항들의 대다수가 신체진찰에 관련된 것이었다. 특히 학생 수준에서 수행하기 어려운 신체진찰 행위인 경우, 대부분 학생들은 0점 또는 2점을 받아서 중간능력을 가진 학생을 선별하는 데 한계가 있음을 발견할 수 있었다. 이는 신체진찰 영역이 다른 어떤 영역보다 부분점수를 부여하는 것이 필요할 것이라는 기존 가정에 의문점을 갖게 한다. 그러나 이것이 부분점수가 타당하지 않거나 필요치 않다는 것을 의미하는 것은 아니다. 다만, 기존 출제문제를 검토할 때 다분문항분석 결과를 활용하는 것이 채점기준표의 점검이 필요한 문항을 선별하는 데 도움을 줄 수 있다는 것이다. 즉, CPX 문항수정 개발 워크숍과 같은 사후 점검을 시행할 때 내용타당성 검토와 더불어 보조적 자료로써 다분문항분석 결과를 활용할 수 있을 것으로 생각된다.

진료수행시험은 문항개발에 많은 인적 물적 자원이 투입되어야 하기 때문에 기출문항을 분석하고 수정 보완하는 과정을 통하여 문항개발의 경험을 축적하는 것이 필요하다. 또한, 다른 시험과는 달리 장기간에 걸쳐 시험이 시행되므로 수험자들 사이의 정보공유로 인한 시험의 안정성이 위협받지 않기 위해서는 난이도, 변별도, 문항추측도가 동등한 다수의 문항을 확보할 수 있는 문제은행 구축이 필요하다. 문제은행구축을 위해서는 문항분석이 필수적으로 이루어져야 한다. 진료수행시험의 문항분석을 고전검사이론으로 하는 것이 더 적합 할 것인지, 아니면 문항반응이론을 적용하는 것이 더 합당할 것인지는 향후 더 많은 연구가 이루어져야 할 것이다. 문항반응이론의 경우, 검사동등화 (Huh, 2005)와 수행형 문항을 채점하는 채점자들의 일관성 정도를 살펴볼 수 있다는 장점 (Park, 2001)이 있다.

결론적으로 2005년 CPX시험을 문항반응이론과 고전검사이론을 이용하여 분석한 결과 전반적으로 중간수준이나 쉬운 문항들로 구성되어 있었으며 변

별도는 고전검사이론과 문항반응이론의 분석 결과가 일치하지 않았다. 다분문항반응이론을 이용한 분석은 부분점수채택 문항의 특성파악과 채점기준표 점검이 필요한 문항을 선별하는 데 보조적 자료로 활용할 수 가능성을 보여주었지만 임상수행평가의 특성상 신중히 적용, 해석되어야 할 것이다. 향후 임상수행평가 문항의 질적 관리와 문제은행을 구축하는데 적절한 문항분석방법에 대한 후속 연구가 지속적으로 이루어져야 할 것이다.

감사의 글

이 연구를 수행할 수 있도록 자료를 제공해주신 2005년도 서울, 경기 지역 CPX 컨소시엄 관계자 및 참여대학에 감사드립니다.

참 고 문 헌

Munro, B. H.(2001). *Statistical Methods for Health Care Research.* Philadelphia: Lippincott Williams & Wilkins.

Huh, S.(2005). Test equating of a medical school lecture examination based on item response theory: a case study. *Korean Journal of Medical Education, 17(1),* 15-25.

Lee, J.S.(1992). Item analysis method according to classical test theory and item response theory. *Yonsei Kyoyook Kwahak, 41,* 18-60.

Lee, Y.M., So, Y.H., Ahn, D.S., Rhee, K.J., Im, H.(2002). Psychometric analysis of comprehensive basic medical science examination. *Korean Journal of Medical Education, 14(2),* 301-306.

Lee, Y.W.(1993). Comparison of item analysis results according to classical test theory and item response theory. *Journal of Education Evaluation, 6(2),* 217-239.

Lim, E.Y., Park, H.H., Kwon, I., Song, G.L., Huh, S.(2004). Comparison of item analysis results of Korean medical licensing examination according to classical test theory and item response theory. *Journal of Educational Evaluation for Health Professions, 1(1),* 67-76.

Park, C.(2001)a. An application of item response theory to an analysis of performance-based items. *Journal of Education Evaluation, 39(2),* 215-232.

Park, C.(2001)b. *Polytomous item response theory.* Seoul: Kyoyook Gwahak Sa.

Seong, T.J.(2005). *Education evaluation.* Seoul: Hakjisa.

Hwang, JK.(1999). *School learning and evaluation.* Seoul: Kyoyook Gwahak Sa.