

데 목적이 있다. 고전검사이론에 따른 문항분석은 이해도 쉽고 분석방법도 현장에 적용하기 쉬우며, 이미 대부분의 평가를 이 방법으로 하여 매우 익숙한 내용이다. 문항을 맞춘 총점으로 수험생 능력을 추정한다. 그러나, 고전검사이론에서는 문항의 난이도, 분별도가 집단의 특성에 따라 달라지며, 수험생의 능력을 시험의 특성에 따라 다르게 추정한다는 단점이 있다. 이러한 문제를 극복하기 위하여 나온 문항반응이론(item response theory)에 따른 문항분석을 시도하면, 각 문항은 문항의 고유한 특성이 있어 수험생 집단의 수준에 따라 문항의 난이도 및 분별도가 달라지지 않고, 수험생 능력을 추정할 때 시험의 난이도에 관계없이 수험생의 능력을 추정할 수 있다[1,2]. 이런 방법은 실제 상황을 더 잘 반영시켜 줄 수 있고 수험생의 능력의 서열을 더 잘 매길 수 있으나, 과거에는 계산방법이 복잡하여 셸틀 플그림(computer program)을 활용하여야만 하기에 흔히 쓰기에는 어려웠다. 최근에는 셸틀(computer)을 문항분석에 쉽게 사용할 수 있고 적절한 플그림(program)도 등장하여 이 문항반응이론에 따른 분석이 쉬어졌다. 그러므로, 문항반응이론에 따라 새롭게 문항 분석하여 난이도, 분별도 값을 구하여, 결과를 해석하면 조금 더 향상된 분석을 할 수 있을 것이다.

문항반응이론에 근거한 문항분석은 우리나라 보건의료 분야의 국가시험에서는 적용한 적이 없다. 실제 시험의 문항분석에 이러한 이론에 따른 방법을 적용하려면 우선 같은 시험 결과를 가지고 새로운 이론에 따른 방법을 적용했을 때 기존의 결과와 얼마나 상관관계가 있는 지 또는 어떤 차이가 있는 지를 파악할 필요가 있다. 그러므로, 최근 시행한 의사국가시험의 결과 자료를 기존의 고전검사이론에 따른 분석 및 문항반응이론에 따른 분석을 적용하여 결과를 비교해 봄으로써 새로운 분석 방법의 활용 가능성 및 그에 따르는 타당성을 검토하려고 하였다. 구체적으로는 두 가지 문항분석 방법에 따른 난이도, 분별도, 추측도와 같은 문항모수 및 능력모수의 상관관계를 살펴보고, 문항반응이론에 따른 문항분석이 의사국가시험 결과 해석에 어떠한 장점이 있는 지를 알아보려고 하였다. 이러한 새로운 문항분석 기법을 도입하여 국시원의 시험 결과 자료를 분석하여 어떻

게 응용할 수 있을지 알아보는 것은 앞으로 문항의 질 향상 및 수험생의 능력 평가 판단에 도움을 줄 수 있을 것이다. 또한 앞으로 셸틀이용 적응검사(computerized adaptive test, 이하 적응검사) 도입의 타당성 평가에 기본 자료를 제공할 수 있을 것이다.

재료 및 방법

2000년과 2001년에 시행된 의사국가시험 제 64회와 65회의 자료를 분석하였다. 제 64회의 경우 7과목(내과학, 외과학, 산부인과학, 소아과학, 정신과학, 예방의학, 보건법규)에 450문항으로 이루어졌고, 응시자가 2,961명인 자료를 사용하였다. 제 65회 자료는 7과목 450문항으로 이루어졌고 3,262명의 자료를 사용하였다. 일차원(unidimensionality) 가정은 문항반응이론 도입 시 전제조건인 하나이다. 일차원 가정은 시험은 응답자의 하나의 능력만을 측정할 수 있어야 한다는 것이다. 일차원 가정을 검증하기 위해서 Hambleton의 검증법을 이용하였다[3]. 요인분석 후 첫 번째 고유 값과 두 번째 고유 값의 차이를 확인하고 스크리 플롯의 확연히 구별되는 지점을 확인하였다. 일차원성 가정은 제 65회 의사국가시험 자료만 검증하였다. 제 65회 의사국가시험을 7개 과목별로 요인 분석하여 각 과목별 일차원성 가정을 검증하였다. 제 64회의 경우는 일차원성 검증은 하지 않았고 과목별로 나누어 문항 분석하여 문항 모수 값을 구하였다. 요인분석을 위해서는 SAS 6.0[4]을 사용하여 분석을 하였다. 문항반응이론 도입 시 또 다른 전제 조건인 문항독립(local independence) 가정은 수험생이 문제를 풀 때 다른 문제를 푸는 데 영향을 미쳐서는 안 된다는 것이다. 즉 수험생이 문항을 맞출 확률은 상호 독립적이어야 한다는 것이다. 문항에서 다른 문항의 정답을 암시하는 내용이 없다고 판단하여 가정을 만족시킨다고 본다.

고전검사이론에 따른 문항모수는 국시원 전산실에서 계산하였다. 분별도는 양극단반분별도로 구하였다. 문항반응이론에 따른 문항분석을 위해서 Analyst 4.0, Bilog 3.11 및 Xcalibre 등 3 가지를 사용하였다[5,6,7]. 각 플그림에서 초기값은 기본(default)으로 설정되어있

는 값을 사용하였다. 문항반응이론에서는 3모수 모델을 사용하였다. 문항모수 중 난이도 및 분별도는 고전검사이론과 문항반응이론에 따른 해를 비교 가능하나, 추측도는 고전검사이론에서는 구할 수 없는 문항모수이므로 비교할 수 없었다. 능력모수 값은 64회와 65회 시험을 과목별로 구분하여 분석하였다. 사용된 프로그램은 Analyst 4.0 및 Xcalibre에서 구해진 능력모수만을 비교하였다. Bilog같은 경우 문항분석 시에 난이도에서 표준오차를 계산하지 못하는 오류가 발생하였고 이 때문에 능력모수가 구해지지 않는 경우가 발생해서 사용하지 못했다. 고전검사이론 및 문항반응이론에 따른 각 모수 값의 피어슨모수상관분석 및 다중상관분석은 SAS 6.0으로 구하였다[4].

결 과

65회 의사국가시험의 과목들을 일차원 가정을 검증하기 위한 요인분석을 시행하여 고유값의 차이 변화를 비교하였다(Table 1).

요인분석을 하면 고유값은 문항 수만큼 계산된다. 그러나 결과 보고서에는 제 1요인, 제 2요인, 제 3요인의 고유값과 제 1, 2, 3, 4, 요인들의 차이 값만 정리하였다. 제 1요인에서 제 2요인간의 고유값 차이가 다른 요인들의 차이보다 확연함을 알 수 있다. 내과의 경우, 제1요인과 제2요인의 차이가 7.7325이고, 제2요인과 제3요인차이는 0.7307, 제3요인과제 4요인의 차이는 0.2525이다. 제 1요인과 제2요인의 차이는 제2요인과 제3요인의 차이보다 7 만크이 더 차이가 난다. 제3요인과 제4요인의 고유값 차이는 0.2525의 2요인과 3요인 고유값 차이인 0.7307에 비해 0.5만크 차이가 난다. 결국 차이를 비교 해보면 제1요인과 제 2요인의 차이는 다른 요인간의 차이에 비해 큰 차이를 보인다. 이렇게 과목별로 요인 분석한 결과, 각 과목은 일차원 가정을 충족하고 있다고 추정할 수 있었다.

국시원에서 고전검사이론에 따라 분석한 의사국가시험의 난이도 및 분별도의 평균과 표준편차를 정리하여 보면, 내과, 외과, 소아과, 산부인과에서 64회가 65회보다는 쉬웠고, 내과, 외과, 소아과, 산부인과, 정신과에

Table 1. Unique value by factor analysis of the 65th Korean Medical Licensing Examination

Class		Factor		
		1st	2nd	3rd
Internal Medicine	Unique value	9.6624	1.9299	1.1992
	difference of unique values	7.7325 ^a	0.7307 ^b	0.2525 ^c
Surgery	Unique value	3.6182	0.7777	0.5929
	difference of unique values	2.8405 ^a	0.1848 ^b	0.1217 ^c
Pediatrics	Unique value	4.1682	1.7349	1.5002
	difference of unique values	2.4333 ^a	0.2347 ^b	0.1177 ^c
Ob & Gy	Unique value	4.5665	0.7034	0.5636
	difference of unique values	3.8631 ^a	0.1398 ^b	0.1010 ^c
Psychiatry	Unique value	1.2893	0.3830	0.2198
	difference of unique values	0.9062 ^a	0.1632 ^b	0.0163 ^c
Preventive medicine	Unique value	2.2417	0.4626	0.3726
	difference of unique values	1.7790 ^a	0.0901 ^b	0.0808 ^c
Medical Law	Unique value	0.6942	0.2349	0.1834
	difference of unique values	0.4593 ^a	0.0516 ^b	0.0595 ^c

a: Difference between unique value of the first factor and that of second factor.

b: Difference between unique value of the second factor and that of third factor.

c: Difference between unique value of the third factor and that of the fourth factor.

서 64회 시험이 65회 시험보다 더 분별력이 있었다 (Table 2).

문항반응이론에 의한 분석 결과에서 문항 난이도 평균은 모두 음수에 속한다 (Table 3).

즉, 이 시험 문제는 대체로 쉽다는 것을 알 수 있다. 고전검사이론의 난이도 및 문항반응이론에 따른 세 가지 풀그림의 난이도 사이의 상관관계에서 제 64회 시험에서 Analyst 4.0과 Xcalibre의 역상관계수가 -0.75로 제일 작았고, 나머지는 모두 -0.75 보다 낮은 값으로 높은 역상관계수를 보인다 (Table 4).

이것은 고전검사이론에서는 난이도가 높은 것이 쉬

운 문제이나, 문항반응이론에서는 낮은 것이 쉬운 문제라는 계산 방법 때문이다. 제 64회에서 고전검사이론에 따른 분별도와 Analyst 4.0 에서의 분별도가 상관계수 0.689로 가장 높은 상관을 보이고 있다. 문항반응이론에 따른 문항분석 풀그림 중에는 Bilog와 Xcalibre에 의한 값이 0.684의 상관을 보이고 있다. 제 65회 경우도 고전검사이론에 따른 분별도와 Analyst 4.0 에서의 분별도가 가장 높은 상관(0.753)을 보이고, 문항분석 풀그림 중에는 Xcalibre와 Bilog의 분별도 상관이 높은 값 (0.55)을 보이고 있다 (Table 5). 추측도를 비교하면, 제 65회 시험에서 Bilog와 Xcalibre로 분석한 문항추측

Table 2. Mean and standard deviation (SD) of difficulty and discriminating index according to classical test theory

Class	64th Exam		65th EXAm	
	Difficulty index Mean(SD)	Discriminating index Mean(SD)	Difficulty index Mean(SD)	Discriminating index Mean(SD)
Total	71.90(21.71)	0.21(0.11)	68.30(23.58)	0.18(0.13)
Internal Medicine	73.47(19.45)	0.23(0.11)	65.19(24.31)	0.21(0.13)
Surgery	73.84(22.50)	0.20(0.100)	69.47(22.03)	0.19(0.11)
Pediatrics	70.62(22.57)	0.23(0.12)	68.39(23.62)	0.21(0.14)
Ob & Gy	73.82(18.84)	0.22(0.11)	65.12(22.53)	0.17(0.12)
Psychiatry	68.63(27.32)	0.17(0.11)	73.58(26.45)	0.14(0.11)
Preventive medicine	65.80(24.98)	0.15(0.10)	73.61(21.56)	0.17(0.13)
Medical Law	68.74(25.97)	0.14(0.09)	68.30(23.58)	0.19(0.13)

Table 3. Mean and standard deviation (SD) of the difficulty parameter, discriminating parameter and guessing parameter analyzed according to item response theory

EXAm	Program	Analyst 4.0 mean(SD)	Bilog mean(SD)	Xcalibre mean(SD)
64th	difficulty	-0.28(1.74)	-1.03(2.70)	-1.0(1.95)
	discriminating	1.18(0.61)	0.49(0.22)	0.46(0.17)
	guessing	0.50(0.02)	0.30(0.07)	0.23(0.05)
65th	difficulty	-0.08(1.85)	-1.11(2.10)	-0.76(2.03)
	discriminating	1.20(0.66)	0.49(0.29)	0.47(0.18)
	guessing	0.50(0.02)	0.21(0.53)	0.21(0.04)

도가 0.347로 가장 높을 뿐 나머지는 상관이 낮다 (Table 6).

64회와 65회 과목별 총점 및 각 분석방법에 따른 능력모수는 제 65회의 의료법규에서 Xcalibre 에서의 값

이 총점 및 Analyst 4.0 에 따른 값과 각각 0.872, 0.897 이라는 상관관계를 보인 것 이외 모두 0.90 이상의 높은 상관을 보인다 (Table 7).

Table 4. Correlation coefficient of difficulty index values of the 64th and 65th National Medical Licensing Examination analyzed by 4 kinds of programs

Exam		Classical test theory	Analyst 4.0	Bilog	Xcalibre
64th	Classical test theory	1.000			
	Analyst 4.0	-0.889	1.000		
	Bilog	-0.752	0.750	1.000	
	Xcalibre	-0.963	0.889	0.787	1.000
65th	Classical test theory	1.000			
	Analyst 4.0	-0.846	1.000		
	Bilog	-0.881	0.880	1.000	
	Xcalibre	-0.960	0.851	0.937	1.000

Table 5. Correlation coefficient of discriminating index values of the 64th and 65th National Medical licensing Examination analyzed by 4 kinds of programs

Exam		Classical test theory	Analyst 4.0	Bilog	Xcalibre
64th	Classical test theory	1.000			
	Analyst 4.0	0.689	1.000		
	Bilog	0.411	0.545	1.000	
	Xcalibre	0.044	0.222	0.684	1.000
65th	Classical test theory	1.000			
	Analyst 4.0	0.753	1.000		
	Bilog	0.227	0.353	1.000	
	Xcalibre	0.023	0.153	0.550	1.000

Table 6. Correlation coefficient of guessing index values of the 64th and 65th National Medical Licensing Examination analyzed by 3 kinds of programs for item response theory

Exam		Analyst 4.0	Bilog	Xcalibre
64th	Analyst 4.0	1.000		
	Bilog	0.097	1.000	
	Xcalibre	0.046	0.164	1.000
65th	Analyst 4.0	1.000		
	Bilog	0.064	1.000	
	Xcalibre	0.208	0.347	1.000

Table 7. Correlation coefficient of total score and ability parameter values of the 64th and 65th National Medical Licensing Examination analyzed by 2 kinds of programs for item response theory

Exam			Total score	Analyst 4.0	Xcalibre	
64th	Internal Medicine	Total score	1.000			
		Analyst 4.0	0.987	1.000		
		Xcalibre	0.982	0.989	1.000	
	Surgery	Total score	1.000			
		Analyst 4.0	0.978	1.000		
		Xcalibre	0.962	0.960	1.000	
	Pediatrics	Total score	1.000			
		Analyst 4.0	0.982	1.000		
		Xcalibre	0.974	0.979	1.000	
	Ob & Gy	Total score	1.000			
		Analyst 4.0	0.975	1.000		
		Xcalibre	0.968	0.977	1.000	
Psychiatry	Total score	1.000				
	Analyst 4.0	0.974	1.000			
	Xcalibre	0.944	0.917	1.000		
Preventive Medicine	Total score	1.000				
	Analyst 4.0	0.967	1.000			
	Xcalibre	0.951	0.920	1.000		
Medical Law	Total score	1.000				
	Analyst 4.0	0.962	1.000			
	Xcalibre	0.943	0.903	1.000		
65th	Internal Medicine	Total score	1.000			
		Analyst 4.0	0.990	1.000		
		Xcalibre	0.987	0.993	1.000	
	Surgery	Total score	1.000			
		Analyst 4.0	0.977	1.000		
		Xcalibre	0.972	0.965	1.000	
	Ob & Gy	Total score	1.000			
		Analyst 4.0	0.972	1.000		
		Xcalibre	0.962	0.959	1.000	
	Pediatrics	Total score	1.000			
		Analyst 4.0	0.980	1.000		
		Xcalibre	0.972	0.976	1.000	
Psychiatry	Total score	1.000				
	Analyst 4.0	0.972	1.000			
	Xcalibre	0.962	0.959	1.000		
Preventive Medicine	Total score	1.000				
	Analyst 4.0	0.973	1.000			
	Xcalibre	0.954	0.952	1.000		
Medical Law	Total score	1.000				
	Analyst 4.0	0.957	1.000			
	Xcalibre	0.872	0.897	1.000		

고찰

이러한 결과는 국시원에서는 문항반응이론에 따른 문항분석을 처음 시도하여 고전검사이론의 해와 비교하여 본 것으로 그 의의를 찾을 수 있다. 문항분석에 앞서, 일차원 검정에 따라 의사 국가시험은 각 과목이 각 과목에 대한 해결 능력을 평가하는 데 적절한 문항으로 구성하였음을 알 수 있었다. 이런 요인분석 역시 의사 국가시험에서는 최초로 시도한 내용으로 국시 문제의 적절성을 잘 설명하여 준다. 상관관계 비교 결과 난이도는 대부분 -0.75 이하라는 큰 역 상관 값을 보여서, 고전검사이론에서의 값과 문항반응이론에서의 값이 역 상관관계가 높음을 알 수 있었다. 즉, 두 방법의 용도가 비록 틀리다 하더라도, 문항반응이론에 따른 분석이 고전검사이론에 따른 분석을 대치하여도 큰 문제가 없음을 알 수 있다. 또한 문항반응이론에 따른 난이도 계산에서 각각 다른 풀그림에서 계산한 값도 서로 상관이 높아서 어느 풀그림을 사용하여도 좋음을 알 수 있다. 분별도는 Analyst 4.0 에 의한 해는 고전검사이론과 상관이 높으나 다른 풀그림에 의한 해는 낮았다. 앞으로 고전검사이론에서도 문항총점상관계수로 구하면 다른 결과가 나올 수 있을 것이다. 추측도는 고전검사이론에서는 구하지 못하고 문항반응이론에서만 구할 수 있는 것으로, 각 풀그림마다의 값이 상관이 매우 낮았는데, 이것은 풀그림의 알고리즘 및 초기 값의 차이에 의한 것으로 추정한다. 능력모수의 총점과 상관관계는 매우 중요한데, 그 이유는 문항반응이론에 따른 능력모수가 높은 학생이 고전검사이론에 따른 총점이 역시 높은 지가 앞으로 문항반응이론의 도입과 여러 분야 적용에 가장 중요한 선결 조건이 될 것이다. 즉, 국가시험과 같은 표준평가에서는 최소한의 점수를 정하는 것이 최소한의 능력을 보유한 것과 같다고 설명할 수 있어야 하기 때문이다. 이런 설명이 가능하다면 문항반응이론으로 문항 분석하여 능력모수를 추정하는 것을 수험생이나 일선 학교 교육자가 쉽게 받아들일 수 있을 것이다. 이 결과에서는 상관관계가 대부분 0.9 이상으로 능력모수로 총점 결과를 대치하여도 무방하다고 볼 수 있다. 이런 능력 모수를 총점에 대치하여 활용하려면 앞으로 능력모수를 기준

점수에 변환하는 작업이 필요하다.

기존의 모형(simulation) 자료가 아닌 실제 자료에서 고전검사이론과 문항반응이론에 따른 문항모수의 상관 관계를 비교한 자료는 많지 않다. 국내에서는 한 대학의 입학학력고사 43 과목에서 두 가지 방법에 따른 문항모수를 비교하여 난이도는 상관계수 평균이 -0.88 로 높은 역상관을 보이고, 분별도는 0.86 으로 역시 높은 상관을 보였다[8]. 미국에서 대규모 주 단위 시험에서는 두 방법에 따른 문항모수 및 능력모수가 매우 유사하다고 하였다[9]. 스웨덴 대학수학능력검사(Swedish Scholastic Ability Test)에서 1996년도 사전시험과 1997년도 정규 시험의 독해 부문 16 문항을 분석한 자료에서 두 방법에 따른 난이도의 상관계수는 각각 -0.9 , -0.92 분별도의 상관계수는 각각 0.35 , 0.78 이었다. 이 결과에서는 비록 두 분석 방법이 다른 이론 배경 때문에 결과가 다르게 나오는 것으로 추론하는 것이 합리적이지만 이런 국가시험과 같은 대단위 시험에서는 두 분석 방법이 상관관계가 크므로, 문항반응이론에서 장점이라고 제시하는 문항모수의 불변성이나 그 외 다른 장점을 찾기 어렵다고 하였다[10].

그렇다면 이런 문항반응이론에 따른 새로운 분석 방법의 활용도는 어디에 있을까? 앞으로 국시원이 적응검사를 의사국가시험에 도입할 수 있다면 문항반응이론에 따른 문항분석 결과를 문항 데이터베이스 구축에 필수 자료로 사용하여야 한다. 문항마다 별개의 특성이 있다는 것이 문항반응이론의 기본 가정이므로, 문항을 데이터베이스에 입력할 때, 난이도 및 분별도를 같이 넣어야 한다. 적응검사에서는 수험생마다 각각 다른 문항을 제시하여 맞춤형(tailored test) 형식으로 수험생이 능력에 적절한 난이도 및 분별도를 가진 문항을 제시하는 방법으로 시험을 구성하므로, 지금처럼 모든 수험생이 같은 문항을 주어진 시간에 치르지 않게 된다. 적응검사에서의 난이도 및 분별도는 고전검사이론에 따라 계산하는 것이 아니라 문항반응이론에 따른 문항분석에서 산출한 값을 사용한다. 이러한 적응검사의 타당성은 더 검토가 필요할 것으로 보이고 장단점에 대한 토의도 필요하다. 그 타당성에 대한 토의 중 하나가 적응검사 도입 시 필요한 문항반응이론에 따른 문항분석에 대한 평가

이다. 이런 새로운 문항분석을 적용하려면 같은 시험 결과를 가지고 고전검사이론 및 문항반응이론에 따른 문항분석을 적용하였을 때, 결과가 어떻게 차이가 나는지 그리고 그 차이 중에서도 어떤 측면의 차이가 어느 정도 되는지를 파악하여야 한다. 위의 결과가 이런 질문에 답을 하고 있다. 즉, 문항반응이론에 따른 수험생의 능력모수가 수험생의 총점과 0.9 이상의 상관관계가 있다는 점에 비추어 문항반응이론에 따른 결과를 총점 대신에 사용하는 것에 대한 논리적인 근거를 제공할 수 있기 때문이다.

이번 연구에서는 다음과 같은 몇 가지 제한점을 제시할 수 있다. 우선 2002년도 제 66회 시험부터는 과목별이 아닌 총론, 각론, 보건법규로 과목이 바뀌었으므로 일차원성 검정을 현재 과목이 아닌, 총론, 각론, 보건법규로 나누어 다시 하여야 하기에, 이 경우 일차원성 가정이 어떻게 변하는 지 파악하여야 한다. 이 연구에서 제시하지 않았지만, 제 65회 시험을 30개 대항목별로 구분하였을 때 역시 일차원 가정을 만족하는 것으로 나타났기에 앞으로도 일차원 가정은 적용에 문제가 없을 것으로 보인다. 또한, 어떤 범주에 시험 문항이 포함되느냐에 따라 문항 모수값이 변할 것임을 알 수 있었다(자료 미 제시). 즉, 과목이 바뀌어도 일차원 가정이 성립하여 문항반응이론에 따른 문항분석이 가능하다.

두 번째로는 문항반응이론에 입각한 문항분석 방법은 수험생 능력에 상관없이 문항의 난이도, 분별도가 추정되는 것이다. 특히 이번 연구의 수험생은 모두 의대생이거나 의대 졸업생으로 이들의 능력 분포가 다양할 것이라고 예상하지 않는다. 그러므로, 이 결과 해석은 상대적으로 비슷한 능력의 집단을 대상으로 하였다는 해석상의 한계가 있다. 일정한 능력을 가진 사람들은 능력에 따라 검사의 난이도와 분별도가 바뀐다는 고전검사이론의 단점에 가장 적게 영향 받는다. 그런 까닭에 분석에 사용되는 응시자들의 수가 많다면 고전검사 분석과 문항반응이론의 검사에서 나온 값들이 큰 차이를 보이지 않을 것으로 추정할 수 있다. 이 연구에서는 난이도 분석의 결과가 이 증거라 할 수 있다. 그러나 분별도와 추측도의 상관분석 결과 상관이 거의 없는 것으로 나타나고 있다. 동질 집단이기에 총점과 능력모수 사이의

상관관계가 매우 높게 나왔을 가능성이 있다.

세 번째로는, 풀그림마다 다른 값을 제시한 점이다. 문항반응이론에 따른 문항분석을 하는 풀그림은 우리나라에서 개발한 Analyst 4.0을 제외하고는 이미 온 누리에서 흔히 사용하는 것이다. 그러나 풀그림마다 난이도 값을 제외하고 분별도, 추측도에서는 상관이 낮았다. 이런 까닭에 난이도를 제외하고 문항분석의 분별도와 추측도는 어떤 풀그림을 사용하는 지에 따라 값이 달라지게 되어 해석에 신중을 기하고, 어느 풀그림을 사용할지는 풀그림의 기본값과 알고리즘을 이해하여 선택하여야 한다.

이런 일의 후속 작업으로 첫째, 7 개 과목별이 아닌 2002년부터 시행한 3 개의 과목별 요인분석과, 문항분석이 필요하다. 둘째, 문항반응이론에서 일모수(난이도), 이모수(난이도 및 분별도) 삼모수(난이도, 분별도 및 추측도) 분석 시의 난이도 및 능력모수의 상관을 분석하여 어느 것을 선택하는 것이 효율이 있는 지를 파악하는 것도 필요하다. 이미 적응검사를 시행하고 있는 캐나다의 의사국가시험은 이모수 분석을 적용하고 미국간호사국가시험에서 일모수 분석을 시행하고 있다. 셋째, 이런 분석은 앞으로 국시원이 적응검사를 도입하고, 기출 문제의 난이도 및 분별도를 문제은행에 추가하여 새로운 시험 문제를 만들어 나가야 그 효용도가 있으므로 이런 적응검사를 국가시험에 도입하는 것에 대한 타당도 조사가 이루어져야 한다. 적응검사를 시행하게 되면 문항반응이론에 따른 분석을 맞춤검사 뿐 아니라 동등화(equating) 검정에 활용할 수 있어, 매년 합격자 수가 문항의 난이도에 따라 변이가 큰 문제를 해결할 수 있을 것이다.

결 론

의사국가시험 결과를 고전검사이론 및 문항반응이론에 따라 문항분석하여 본 결과, 문항반응이론의 난이도 및 능력모수는 고전검사이론의 난이도 및 총점과 각각 높은 상관이 있음을 알 수 있었다. 국가시험과 같은 자격시험에서는 능력모수 및 총점이 합격, 또는 불합격을 판가름하는데 가장 중요하므로, 이러한 결과는 앞으

로 국시원에서 문항반응이론을 활용한 적응검사 도입의 타당도를 뒷받침한다. 앞으로 의사국가시험 뿐 아니라 여러 직종의 시험 결과를 재료로 하여 다양한 문항반응이론의 적용에 대하여 연구할 필요가 있다고 제안한다.

Measurement, 1999 (available from http://www.umu.se/edmeas/publikationer/index_eng.html).

핵심
단어

교육평가, 문항반응이론, 고전검사이론

참고 문헌

1. 성태제. 문항반응이론 입문. 서울: 양서원, 1991.
2. 성태제. 문항반응이론의 이해와 적용. 서울: 교육과학사, 2001
3. Hambleton RK, Swaminathan H. Item response theory: Principles and applications. Boston Kluwer Academic Publishers, 1985.
4. SAS [Computer program] version 6.0. North Carolina, U.S.A.: SAS Institute Inc., 1995.
5. Analyst [Computer program] Version 4.0 Chunchon, Korea: Hallym University Institute of Medical Education, <http://analyst.hallym.ac.kr>
6. Bilog [Computer program] version 3.11. Illinois, U.S.A.: SSSI Scientific software, 1990.
7. Xcalibre [Computer program] Mineapolis, U.S.A.: Assessment System Corporation, 1999.
8. 이연우. 고전검사이론과 문항반응이론의 문항통계치의 비교. 교육평가연구. 1993; 6(2): 217-239.
9. Fan XT. Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational Psychological Measurement, 1998; 58(3): 357-381.
10. Stage C. A comparison of item analysis based on item response theory and classical test theory. A study of SweSAT subset READ. In: Educational Measurement No. 33. Umea University Department of Educational

