大韓本草學會誌 제37권 제2호(2022년 3월)
Kor. J. Herbol. 2022 ; 37(2) : 13-21

ISSN 1229-1765(Print), ISSN 2288-7199(Online)
http://dx.doi.org/10.6116/kjh.2022.37.2.13.

문항반응이론을 활용한 한의학 교육에서 본초학 시험문항에 대한 연구

채한[1#], 한상윤[2#], 양기영[1], 김형우[1*]

1 : 부산대학교 한의학전문대학원, 2 : 대전대학교 한의학과

Study on the herbology test items in Korean medicine education using Item Response Theory

Han Chae[1#], Sang Yun Han[2#], GiYoung Yang[1], Hyungwoo Kim[1*]

1 : School of Korean Medicine, Pusan National University, Busan, 50610, South Korea
2 : College of Korean Medicine, Daejeon University, Daejeon, 34520, South Korea

ABSTRACT

Objectives : The evaluation of academic achievement is pivotal for establishing accurate direction and adequate level of medical education. The purpose of this study was to firstly establish innovative item analysis technique of Item Response Theory (IRT) for analyzing multiple-choice test of herbology in the traditional Korean medicine education which has not been available for the difficulty of test theory and statistical calculation.

Methods : The answers of 390 students (2012-2018) to the 14 item herbology test in college of Korean medicine were used for the item analysis. As for the multidimensional analysis of item characteristics, difficulty, discrimination, and guessing parameters along with item-total correlation and percentage of correct answer were calculated using Classical Test Theory (CTT) and IRT.

Results : The validity parameters of strong and weak items were illustrated in multiple perspectives. There were 4 items with six acceptable index scores, and 5 items with only one acceptable index score. The item discrimination of IRT was found to have no significant correlation with difficulty and discrimination indices of CTT emphasizing attention of professionals of medical education as for the test credibility.

Conclusion : The critical suggestions for the development, utilization and revision of test items in the e-learning and evidence-based Teaching era were made based on the results of item analysis using IRT. The current study would firstly provide foundation for upgrading the quality of Korean medicine education using test theory.

Key words : Classical Test Theory, e-Learning, herbology, Evidence-based Teaching, Item Response Theory, medical education

*Corresponding author : Hyungwoo Kim, KMD PhD School of Korean Medicine, Pusan National University, Yangsan City, Gyeongnam, 50612, South Korea.
· Tel : +82-51-510-8458 · E-mail : kronos7@pusan.ac.kr
#First author : Han Chae, KMD PhD School of Korean Medicine, Pusan National University, 49, Busandaehak-ro, Mulgeum-eup, Yangsan-si, Gyeongsangnam-do, 50610, South Korea.
· Tel : +82-51-510-8470 · E-mail : han@chaelab.org
Sang Yun Han, KMD PhD College of Korean Medicine, Daejeon University 62 Daehak-ro, Dong-gu, Daejeon, 34520, South Korea.
· Tel : +82-42-280-2634 · E-mail : drhan@dju.kr
· Received : 24 January 2022 · Revised : 14 March 2022 · Accepted : 25 March 2022

Ⅰ. 서 론

의학교육에 있어서 시험 또는 학업성취도 평가는 과거의 학습에 대한 성적을 부여함과 동시에 부족한 부분에 대한 미래의 학습과 새로운 교육의 계획을 위한 기초적인 자료이다[1-3]. 평가의 구조와 내용은 학습이 지향하는 내용과 수준을 결정하기에 가시적인 성적을 넘어 보이지 않는 교육의 방향과 질을 결정하는 교육 현장의 숨겨진 주역이다[4,5].

검사(test)는 체계적인 절차를 통해 내재적 속성을 간접적으로 측정하는 평가 도구로서, 검사 이론(test theory) 또는 문항 분석(item analysis)은 검사와 검사를 구성하는 문항들이 의도한 목적에 부합하는지를 다양한 측면에서 분석한다. 좋은 문항과 검사는 의학 교육, 의료인의 면허 시험, 어학 인증 시험, 수행 평가, 고등학생의 수학 능력, 적성, 인지 능력, 지적 능력, 성격, 자아 개념, 정서적 특성, 스트레스, 정신과 질환에서의 임상 특성 등에서 다양한 잠재적 특성(latent trait)을 측정하기위한 핵심적인 요소이므로[1,5-8], 이들을 분석하기 위한 다양한 타당도 지표가 지속적으로 개발되고 사용되어 왔다[3,9,10].

검사이론은 크게 나누어 고전적검사이론(Classical Test Theory, CTT)과 문항반응이론(Item Response Theory, IRT)이 있다[11,12]. CTT는 측정과 관찰 점수에 대한 찰스 스피어만(Charles Spearman)의 기본 개념을 토대로 검사 총점을 분석의 기본 단위로 사용하는데, 계산 방법이 단순하고 이해와 사용이 용이하기에 19세기 말부터 검사의 기본적인 타당도 척도로 사용되었다. CTT에 있어서 검사 및 문항의 난이도(item difficulty)는 피험자가 맞춘 개수를 기준으로, 개수가 많으면 쉽고 적으면 어렵다고 평가한다. 또한, 문항 타당도(item discrimination)는 능력별 맞춘 횟수의 차이를 기준으로, 우수한 능력자와 부족한 능력자의 검사 및 문항 점수에 차이가 클수록 타당도가 높은 것으로 평가한다.

그러나, CTT는 피험자의 능력과 검사를 함께 시행했던 동료들의 능력, 검사(문항)의 종류와 구성에 따라 난이도와 변별도가 매번 달라지기에 태생적인 상대평가라는 문제를 지니고 있다[9,12,13]. 예를 들어, 피험자가 5문제를 맞추었다고 해도, 5개의 어려운 문제와 5개의 쉬운 문제를 해결하기 위한 능력에는 명백한 차이가 존재한다. 아울러, 동일한 6문제 검사라도 검사를 함께 시행했던 동료들이 5문제를 맞추면 쉬운 시험이 되며, 5문제를 틀리면 어려운 시험이 되는데, 두 경우 모두 타당도가 낮은 것으로 평가된다.

그러나, IRT는 변하지 않는 피험자들의 고유한 능력을 중시하는 잠재적 특성 이론(latent trait theory)을 토대로, 문항 특성 곡선(Item Characteristic Curve, ICC)을 사용하여 각 문항의 고유한 특성을 분석하므로, CTT의 문제점들을 보완하면서 절대평가의 도구로 활발하게 사용되고 있다. ICC는 피험자의 능력(theta, θ)과 문항의 답을 맞힐 확률(probability, $P(\theta)$) 사이의 함수관계를 나타내는데, 일반적으로 Figure 1과 같은 S자 형태의 곡선으로 표현된다[11-13]. IRT에서의 문항 변별도(α)는 답을 맞출 확률(probability)이 절반((1+c)/2)에 해당되는 ICC에서의 기울기이고(Figure 1), 문항 난이도(β)는 ICC에서 답을 맞출 확률이 절반에 해당되는 점에서의 능력(theta) 수준이며, 문항 추측도(item guessing, c)는 정답을 몰라도 맞출 수 있는 확률이다.

IRT를 사용하면 검사의 난이도와 변별도를 일정하게 유지할 수 있으므로, 임상 진단검사의 개발이나 자격 시험용 문제 은행의 관리, 컴퓨터기반 검사(Computer-based Test, CBT)의 개발과 운용에 필수적으로 활용된다[5,7,14]. IRT는 교육용 시험 문항과 임상 검사의 개발에 필수적인 통계분석 기법이지만[4,13,15], 상황과 목적에 따라 관심을 두는 능력 특성에는 차이를 보인다[7]. 일상적인 대학 교육이나 임상 검사에서는 현장에서 만나는 다양한 특성 수준을 측정하기위해 유효 측정 영역을 폭넓게 설정하지만, 자격 시험 및 입학 시험에서는 합격과 불합격을 가르는 특정 범위에서의 타당도를 강조하게 된다[1,16].

이처럼 다양한 검사 현장에서의 핵심적인 중요성에도 불구하고, IRT는 개념의 이해가 어렵고 수학적 분석이나 계산도 매우 복잡하여 일반 연구자들이 용이하게 활용할 수 없었다[4,7,12]. PARSCALE, XCALIBRE, WINSTEPS, BILOG-MG와 같은 전용 소프트웨어나 R 패키지 등은 통계 전문가를 위한 다양한 분석방법을 제공하고는 있으나, IRT 개념의 충분한 이해와 텍스트 기반 명령어에 대한 숙달이 전제되어야 한다[12,17]. 이러한 한계를 극복하기 위한 새로운 소프트웨어들이 개발되고 있는데, 본 연구에서의 jMetrik은 2020년 성정(性情)과 소증(素證)을 측정하는 임상 검사의 개발과 타당화 과정에서 한의학에 처음으로 도입된 문항 및 검사분석 도구이다[13,18-20].

의학교육 분야에 있어서 정보통신기술(Information and Communication Technology, ICT)을 사용하는 교육을 의미하는 이러닝(e-learning)은 빠르게 발전하고 있는데[21-23], 시험 평가에 있어서도 컴퓨터와 태블릿과 같은 정보통신 기기를 사용하는 CBT나 피험자 능력을 고려한 컴퓨터기반 맞춤형 검사(Computerized Adaptive Testing, CAT)의 사용 또한 급속도로 확산되고 있다[7,24]. CBT는 보건의료인 국가시험에도 적극적으로 도입되고 있는데[7], 의사(2022) 직종에 이어 2023년부터 한의사와 치과의사 직종에도 도입될 예정이며[25], 미국 의사 자격시험(United States Medical Licensing Examination)이나 간호사 자격시험(National Council Licensure Examination for Registered Nurse) 등의 전문가 자격시험에는 이미 활용되고 있으며[6,15,26], 한의학과 중간 및 기말 고사에는 2013년부터 활용되고 있다.

이와 같은 CBT 및 CAT 시스템을 기존의 지필 검사와 비교한다면, 의학 교육 현장에서의 시험 출제와 관리가 용이하며, 학생들의 문항별 반응을 실시간으로 확인하며, 시험에 사용되는 문항의 숫자와 시간을 최소한으로 유지하는 논리적 근거를 제공한다[5,14,24]. 학습 관리 시스템(Learning Management System, LMS)과의 적절한 연계를 통해 CBT를 학업 성취도 관리와 향상에 유용하게 활용하기 위해서는, 이들의 기반이 되는 효율적인 문항 개발이 선행되어야 한다[4,22,23,26].

이에, 본 연구에서는 한의학과 학생 390명의 본초학 시험에 8년간 동일하게 사용되었던 14개의 시험 문항을 대상으로 CTT와 IRT에 기반한 문항분석을 시행하였다. 기존의 CTT는 한의사 직역 국가시험과 대학 교육에서의 오랜 경험으로 충분히 이해되고 있으나, 새로운 IRT는 개념부터 낯설고 교육 현장에서의 활용 경험도 없었다.

본 연구에서는 두가지 분석법을 일상 한의학 교육 현장의

시험 문항에 적용되었을 때의 유효성을 검토하였으며[24], IRT의 난이도, 타당도, 추측도 등의 문항 파라미터가 한의학 교육에 활용되기 위한 토대를 마련하고자 하였다. 이와 함께, 문항 타당도의 분석 과정에 보건의료인 국가시험원의 기준을 사용하여 타 직역 의료인과 동일한 기준이 활용될 수 있도록 하였다[4,6,24].

문항분석 및 이를 토대로 한 검사이론은 한의학 임상과 교육 두 가지 모두에 있어서 신뢰성과 타당도를 담보하기위한 핵심적인 통계로, 근거기반교육(Evidence-based Teaching)[27-29]을 위한 필수적인 연구 방법론이다. 본 연구를 통해 미래 ICT 환경에서의 한의학 교육을 지원하기위한 시스템적 토대가 마련될 수 있을 것이다[30].

Ⅱ. 연구 방법

1. 시험 문항 및 연구 대상

본 연구에는 A형 및 K형으로 구성된 동일한 본초학 시험(14문항)에 대한 재학생(2011~2018년) 390명의 응답이 사용되었다. A형 문항은 하나의 문항 줄기에 대한 5개의 답가지로 구성된 것으로, 예를 들어 '○○○의 특징, 기미(氣味), 약성(藥性), 치료 효과로 가장 옳은 것을 고르시오'와 같다[9]. K형 문항은 문항줄기에 해당되는 물음과 함께 4개 항목의 보기를 제시하고, 5개의 답가지에서 문제의 답으로 적절하게 조합된 짝 하나를 고르는 형식으로 구성된 것으로, 예를 들어 '○○○의 다음 네 가지 설명을 대상으로 옳게 조합된 짝을 고르시오'와 같다. 본 연구는 기관 연구윤리위원회의 승인(2021_148_HR) 이후에 데이터를 정리하고 분석하였으며, 피험자를 확인할 수 있는 인적 정보는 분석 과정에 포함되지 않았다.

2. 고전검사이론(Classical Test Theory, CTT)

CTT에서의 문항 난이도는 문제가 쉬운 정도, 정답률 또는 시험을 본 학생들이 해당 문항의 답을 맞힌 비율로, 0.00 ~ 1.00 사이의 범위에 있다. 본 연구에서는 한국보건의료인국가시험원[31]의 해석 방법을 사용하여 0.50 ~ 0.60을 '최적 범위'로, 0.30 ~ 0.70를 '허용 범위'로 분류하였다. 이와 함께, 정답률이 0.70 이상인 문항을 '쉬운 문항'으로, 0.30 이하인 문항을 '어려운 문항'으로 해석하였다[9].

CTT에서의 문항 변별도는 능력이 높은 학생과 낮은 학생을 변별하는 정도 또는 시험에 사용된 모든 문항에 대한 점수를 기준으로 상위(27%) 집단의 해당 문항에 대한 정답률에서 하위(27%) 집단의 정답률을 뺀 값으로[13,31], −1.0 ~ +1.0 사이의 범위에 있다. 본 연구에서는, 0.15 이하를 불량, 0.15 ~ 0.25는 경계, 0.25~0.35는 양호, 0.35 이상을 우수로 해석하는 한국보건의료인 국가시험원의 해석방법을 사용하였다[31].

3. 문항반응이론(Item Response Theory, IRT)

IRT에서의 문항 난이도(β)는 문제의 어려운 정도 또는 해당 문항의 답을 맞출 확률(probability)이 ICC상 중간((1+c)/2)에 해당되는 점에서의 능력(ability 또는 theta) 수준으로(Figure 1), 일반적으로 −2.0 ~ +2.0 사이의 범위에 있다[13]. −2.0 이하이면 난이도가 '매우 쉽다', −2.0 ~ −0.5는 '쉽다', −0.5 ~ 0.5는 '중간이다' 또는 '보통이다', 0.5 ~ 2.0는 '어렵다', 2.0 이상이면 '매우 어렵다'고 해석하였다[12,31].

IRT에서의 문항 변별도(α)는 능력이 낮은 학생과 높은 학생을 변별하는 정도 또는 해당 문항의 답을 맞출 확률이 ICC상 중간에 해당되는 점에서의 기울기로서(Figure 1), 일반적으로 0 ~ +2.0 사이의 범위에 있다. 본 연구에서는, 0.34 이하이면 변별력이 '거의 없다', 0.35 ~ 0.64는 '낮다', 0.65 ~ 1.34는 '적절하다', 1.35 ~ 1.69는 '높다', 1.70 이상이면 '매우 높다'로 해석하였다[12,31].

IRT에서의 문항 추측도(c)는 객관식 문항에서 정답을 모르면서도 맞출 수 있는 값 또는 능력이 제일 낮은 학생이 정답을 맞출 확률로, ICC와 확률 축이 만나는 절편 값을 말한다. 0 ~ 1 사이의 범위에 있으며, 본 연구에서는 0 ~ 0.2 사이를 적절한 영역, 0.2 ~ 0.3 사이를 경계 영역으로, 0.3이상을 과도하게 높은 영역으로 해석하였다[31].

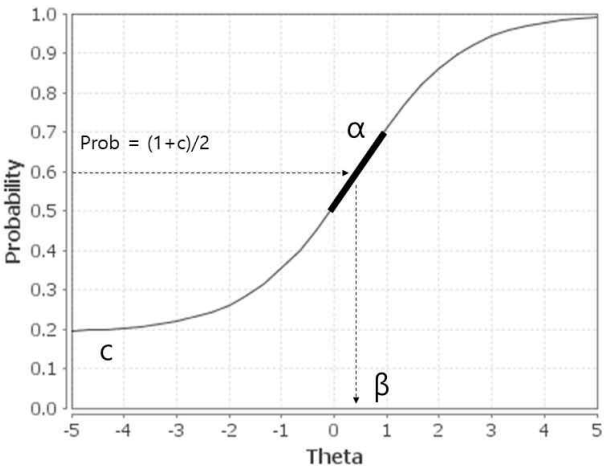

Figure 1. Item Characteristics Curve showing item difficulty (β), discrimination (α) and guessing (c) parameters with ability (theta) and estimated corresponding probability in 3 parameter IRT model.

4. 통계 처리

문항의 타당도를 분석함에 있어서, CTT를 사용한 경우에는 답가지의 선택 빈도, 난이도, 변별도 및 문항점수 − 총점점수의 상관성을 계산하였으며, IRT를 사용한 경우에는 3 모수 로지스틱 모형(3 parameter logistic model)과 주변최대우도(Marginal Maximum Likelihood)법을 사용하여 난이도, 변별도, 추측도를 계산하였다. CTT와 IRT의 난이도와 변별도, 추측도간 상관성 분석에는 피어슨 상관분석과 산점도를 사용하였다. 모든 문항 분석에는 jMetrik 4.1.1(J. Patrick Meyer, Charlottesville, VA)을 사용하였으며[13], 통계적 유의성을 확인하기 위한 p값으로는 0.05, 0.01, 0.001을 사용하였다.

Ⅲ. 결 과

1. 문항 및 문항 응답 특성

문항 및 문항 응답에 대한 분석 결과는 Table 1과 같다. 14개의 선다형 지식역량 암기형 문항에 있어서, A형 문항은 10개, K형 문항은 10개(12, 14, 15, 16번)였다. 문항별 답 가지 선택 빈도를 분석한 결과, 정답의 선택 비율은 52.8%(12번 문항)에서 97.2%(13번 문항)의 범위를 보였다. 12번 문항은 1번(20.3%)과 3번(19.5%) 답가지를 매력적인 오답으로 지니고 있었으며, 10번 문항은 1번(9.2%), 2번(6.9%), 3번(17.2%), 4번(7.4%) 답가지가 매력적인 오답인 것으로 확인되었다. 이와 대조적으로 13번 문항은 정답 답 가지(97.2%) 외의 모든 오답 답가지들이 1.0% 이하의 선택 비율을 보였다.

2. 문항 분석 결과

1) CTT를 사용한 분석 결과

CTT를 사용하여 난이도를 분석한 결과, 4개(7, 9, 10, 12번 문항) 문항이 허용범위(0.30 ~ 0.70)에 속하였으며, 2개(10, 12번 문항) 문항이 최적범위(0.50 ~ 0.60)에 해당되었다. 이와 함께, 어려운(0.30 이하) 문항은 없었으며, 쉬운(0.70 이상) 문항 10개 중에서 4개(2, 4, 6, 11번 문항)가 0.70대의 난이도를 보였다(Table 1).

CTT를 사용하여 변별도를 분석한 결과, 불량(0.15 이하) 문항은 1개(13번), 경계(0.15 ~ 0.25) 문항은 2개(5, 14번), 양호(0.25 ~ 0.35) 문항은 3개(1, 15, 16번), 우수(0.35 이상) 문항은 8개(2, 4, 6, 7, 9, 10, 11, 12번) 이었다. 문항별 점수와 총점(14문항) 간의 상관성을 분석한 결과, 모든 문항들이 총점과 유의한 상관성을 지니고 있는 것으로 확인되었다. 문항점수와 총점 간의 상관계수가 0.4이상인 문항은 모두 7개(4, 7 ~ 12, 16번 문항)로 확인되었다.

2) IRT를 사용한 분석 결과

IRT를 사용하여 난이도를 분석한 결과, 쉬운(−2.0 ~ −0.5) 문항은 10개이었으며, 보통 또는 중간(−0.5 ~ +0.5)의 난이도를 가진 문항은 4개(7, 9, 10, 12번)이었고, 어려운(0.5 ~ 2.0) 문항은 없었다. 문항반응이론을 사용하여 변별도를 분석한 결과, 낮은(0.35 ~ 0.64) 문항은 2개(2, 15번)였으며, 적절한(0.65 ~ 1.34) 문항은 12개였다. 문항반응이론을 사용하여 문항 추측도를 분석한 결과 적절한(0.2 이하) 문항은 2개(7, 12번)였으며, 경계(0.2 ~ 0.3)는 7개였고, 과도한(0.3 이상) 문항은 5개(1, 5, 13, 14, 16번)로 확인되었다.

Table 1. Results of Item analysis using Classic Test Theory (CTT) and Item Response Theory (IRT).

Item	Type	Frequency (%)						CTT#			IRT#		
		1	2	3	4	5	n.r.	DF	DS	CC	DF	DS	GU
item01	A	1.0	1.3	4.1	5.1	87.4*	1.0	0.87	0.27	0.35**	−1.339	**0.955**	0.500
item02	A	7.7	9.5	7.7	71.8*	2.3	1.0	0.72	**0.39**	0.33**	−1.323	0.435	**0.229**
item04	A	1.5	77.2*	5.4	12.8	2.1	1.0	0.77	**0.42**	**0.42****	−1.198	**0.847**	**0.222**
item05	A	0.8	91.0*	3.8	0.5	3.1	0.8	0.91	0.24	0.39**	−1.563	**1.275**	0.500
item06	A	0.8	7.9	10.8	76.7*	2.6	1.3	0.77	**0.46**	0.39**	−1.282	**0.742**	**0.221**
item07	A	60.8*	7.4	5.4	9.5	15.1	1.8	**0.61**	**0.61**	**0.47****	**−0.055**	**1.026**	**0.198**
item09	A	6.4	61.0*	7.7	8.5	14.9	1.5	**0.61**	**0.64**	**0.47****	**−0.056**	**1.155**	**0.201**
item10	A	9.2	6.9	17.2	7.4	55.9*	3.3	**0.56**	**0.52**	**0.40****	**0.373**	**0.721**	**0.213**
item11	A	14.6	1.8	5.1	77.2*	0.5	0.8	0.77	**0.47**	**0.43****	−0.985	**1.114**	**0.221**
item12	K	20.3	5.1	19.5	0.8	52.8*	1.5	**0.53**	**0.61**	**0.44****	**0.409**	**0.988**	**0.192**
item13	A	0.3	0.5	97.2*	0.5	1.0	0.5	0.97	0.07	0.22**	−3.554	**0.898**	0.500
item14	K	1.3	1.8	1.0	1.3	93.8*	0.8	0.94	0.17	0.31**	−2.230	**1.071**	0.500
item15	K	0.3	17.4	0.3	80.8*	0.5	0.8	0.81	0.34	0.31**	−2.142	0.548	**0.229**
item16	K	3.6	3.1	87.9*	4.6	0.0	0.8	0.88	0.33	**0.42****	−1.223	**1.292**	0.500

n.r., no response; DF, Difficulty parameter; DS, Discrimination parameter; CC, Correlation coefficient between item score and total score; GU, Guessing parameter.
*, correct answer; **, p < 0.001; #, Bold represents acceptable (0.30 ~ 0.70) difficulty, good (0 > 0.35) discrimination, and meaningful (0 > 0.04) correlation coefficient in CTT and moderate (−0.5 ~ 0.5) difficulty, adequate (0 > 0.65) discrimination, and acceptable (0 < 0.3) guessing parameter in IRT.

3. 문항 타당도 지표간 상관성

문항분석 타당도 지표들 사이의 상관성 분석 결과는 Table 2 및 Figure 2과 같다. CTT와 IRT 사이의 상관 관계에 있어서, 난이도는 유의하고도 높은 부적 상관관계(r = −0.900, p 〈 0.001)를 보였으며(Table 2), 모두 적절한 것은 4개(7, 9, 10, 12번) 문항이었다(Figure 2−A and Table 1). 또한 두 변별도는 유의하지 않은 상관관계(r = −0.029, p = 0.922)를 보였는데(Table 2), 모두 적절한 것은 7개(4, 6, 7, 9, 10, 11, 12번) 문항이었다(Figure 2 − B and Table 1). CTT에서 난이도와 변별도는 높은 정적 상관관계(r = −0.933, p 〈 0.001)을 보였으며, 두가지 모두 만족한 것은 4개(7, 9, 10, 12번) 문항이었다(Figure 2 − C and Table 1). IRT에서 난이도와 변별도는 유의하지 않은 상관관계(r = 0.111, p = 0.707)을 보였으며, 두가지 모두 만족스러운 것은 4개(7, 9, 10, 12번) 문항이었다(Figure 2 − D and Table 1). IRT의 난이도와 CTT의 변별도(r = 0.910, p 〈 0.001)는 상관성이 유의하였으나, IRT의 변별도와 CTT의 난이도(r = −0.203, p = 0.487)는 유의한 상관관계를 보이지 않았다(Table 2).

IRT의 문항 예측도는 IRT의 문항 변별도(r = 0.458, p = 0.100)와는 유의한 상관성을 보이지 않았으나, IRT의 난이도(r = −0.629, p = 0.016), CTT의 변별도(r = −0.844, p 〈 0.001), CTT의 문항 변별도(r = 0.835, p 〈 0.001)와는 유의한 상관관계를 보였다(Table 2).

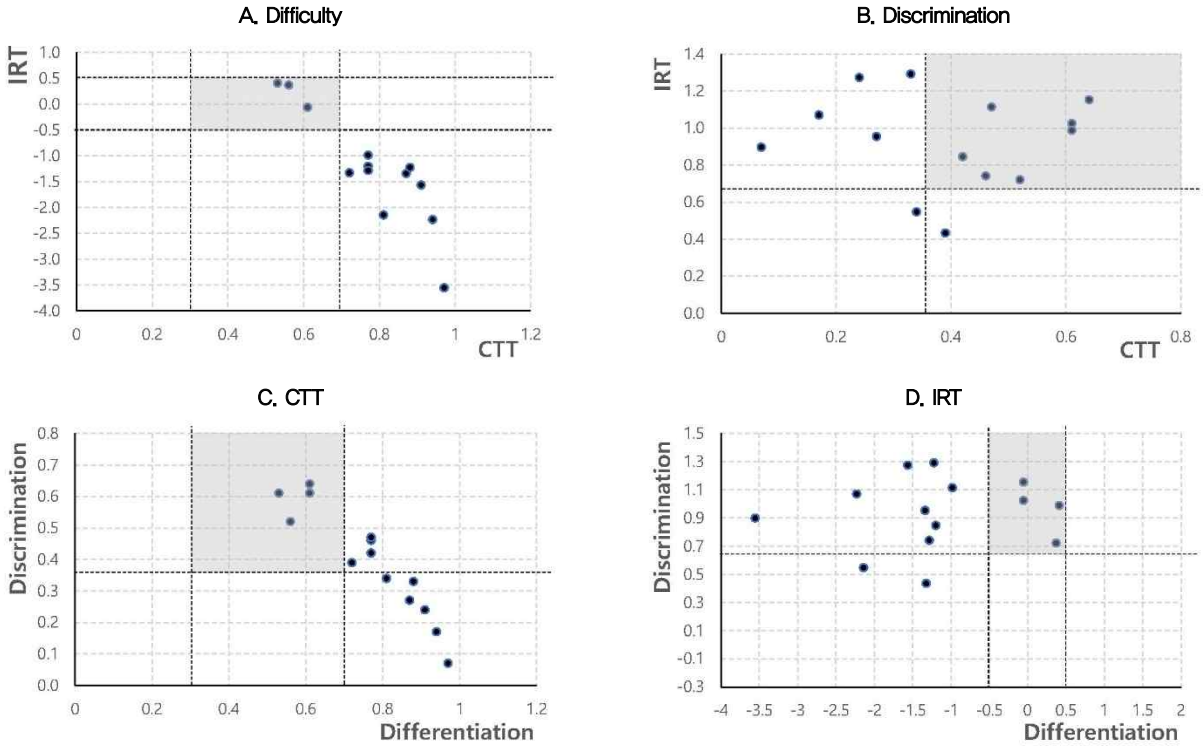

Figure 2. Illustrated correlation between validity measures in IRT and CTT. Shade and dotted lines represent area with acceptable validity parameters. A, Difficulty parameter with IRT and CTT. B, Discrimination parameter with IRT and CTT. C, Discrimination and difficulty parameters using CTT. D, Discrimination and difficulty parameters using IRT

Table 2. Correlation coefficient between validity indices of CTT and IRT.

	Item Response Theory		
	Difficulty	**Discrimination**	**Guessing**
Classical Test Theory			
Difficulty	**−0.900****	−0.203	**0.835****
Discrimination	**0.910*****	−0.029	**−0.844****

**. P 〈 0.001. Bold represent correlation coefficient bigger than 0.4.

IV. 고 찰

본 연구는 한의학과에서 교육을 위하여 사용된 본초학 시험 문항의 다면적 타당도 분석을 진행하였으며[24], 한의학계에서는 처음으로 CTT와 IRT를 동시에 사용하여 난이도, 변별도, 추측도, 문항-총점 상관계수, 정답률을 제시하였다. CTT는 개념도 익숙하고 오랫동안 사용하여 왔기에 문항 개발에 잘 반영되어 왔으나, 이러닝 시대를 위한 IRT는 이해하기 어려우면서도 계산도 난해하여 한의계에 도입되지 못하여 왔다[32]. 본 연구에서는 IRT를 한의약 임상검사에 사용하였던 경험[18,19]을 토대로 시험 문항의 분석에 처음으로 활용하였다.

분석 결과를 통해 얻은 한의학 교육학적으로 중요한 함의들은 다음과 같다. 첫째, 문항 분석 개념을 명시적으로 활용하지는 않았었음에도 불구하고, 학기 중 시험에 만족스럽게 반영되고 있었음을 확인할 수 있었다. 예를 들어, 14 문항 중 4문항(7, 9, 10, 11, 12번)은 모든 타당도 지표를 만족하고 있었는데, 특히 7번(A형)과 12번(K형)은 문제의 유형과 무관하게 높은 타당도를 지니고 있었다(Table 1, Figure 2)[3].

이와 함께, 정답율이 90% 이상이면서 CTT난이도가 0.90 이상이고, IRT 문항 추측도가 과도한(0.3 이상) 문항인 5번, 13번, 14번 문항은 강한 독성으로 극렬한 부작용을 지니거나 소량의 발암물질을 함유하는 한약재에 대한 것으로, 임상 사용에서의 주의를 크게 강조하였던 것이 학생들의 학습에 잘 반영되어 손쉽게 풀었음을 보여준다. IRT 추측도는 능력이 가장 낮은 학생들이 답을 맞출 수 있는 확률로서, 개념적으로 가장 쉬운 문항 또는 CTT 난이도가 가장 높은 문항과 유사하다. 이러한 사실은, IRT 추측도와 CTT 난이도 사이의 상관성이 높은($r = 0.835$, $p < 0.001$) 것이나, IRT 추측도가 과도한(0.3 이상) 5개(1, 5, 13, 14, 16번) 문항들이 CTT 난이도 또한 0.81 이상으로 쉬운(0.7 이상) 범주에 해당하였다는 것으로 확인된다(Table 1 and 2).

본 문항분석 연구의 기획 단계에서는 본초학에서의 단순 지식을 확인하는 문제들이기에 타당도가 매우 낮을 것이라는 회의적인 선입견들을 우려하였었으나, 연구 결과를 통해 한의학 교육 현장에서의 시험 문항들이 본초학 학습목표에서 제시한 지식 역량을 측정하기에 충분한 타당도를 지니도록 제작되고 있음을 재확인할 수 있었다. 이에, 기초학 과목과 함께 임상 교육에서의 다양한 시험 문항들도 한의사 직무 및 교과별 학습 목표에서 제시하는 지식, 기술, 태도 역량들을 평가하기에 적절한 타당도를 지니고 있는지 IRT를 사용하여 다면적으로 분석되어야 할 것이다[10].

둘째, IRT를 사용한 변별도는 다른 타당도 지표와는 유의하지 않은 상관성을 보임으로서 독립적인 특성이라는 것을 확인할 수 있었는데(Table 2, Figure 2), IRT 변별도는 문항특성곡선(Figure 1)에서의 기울기 또는 난이도에 해당하는 역량 특성 수준에서 능력 차이를 구별해내는 능력으로서, CTT 변별도와는 이름만 동일할 뿐 계산 방법이나 결과의 표현, 그리고 그 의미가 전혀 다르다.

CTT 변별도는 상대평가를 기준으로 특정 집단에서의 상대적 순서를 기준으로 하지만, IRT 변별도는 절대평가에서 학생들의 고유한 능력을 구분해내는 능력을 의미하는 것이기에 CBT 및 CAT에서 핵심적인 역할을 수행한다. 이는 향후 시험 문제의 출제 과정에서 IRT 변별도가 적극적으로 고려되어야 함을 의미한다. 특히, 국내외 의약계열의 시험 및 문항 분석에 이미 적극적으로 도입되었으며, 한국 보건의료인 국가시험원에서도 다양한 정책연구를 진행해온 것을 고려한다면[1,2,4,26], 한의계의 움직임이 시급히 요구된다고 할 것이다[33].

본 연구를 통해 IRT를 사용한 문항분석이 한의학 교육학의 중요한 연구도구라는 것이 재확인되었는데, 비록 시작은 늦었지만 적극적인 활용을 통해 한의학 교육의 질을 빠르게 향상시킬 수 있을 것이다[6]. 이에, 한의학 교육기관의 교수자 및 미래 교수자들이 본 연구에서의 IRT를 포함한 다양한 타당도, 신뢰도 분석 기법을 손쉽게 사용할 수 있는 시스템이 구축되어야 할 것이다.

셋째, 시험 문항의 개발에 있어서 특정 문항이 어떠한 수준의 학업역량을 지닌 학생들을 주요한 평가 대상으로 할 것인지 사전에 정밀하게 기획되어야 한다. 본 연구에서의 객관식 시험과 동시에 시행되었던 주관식 문제가 높은 난이도(상위권, 높은 본초학 학업성취도 또는 높은 학업역량을 지닌 학생들)를 목적으로 사용되었음을 고려한다면, 본 연구에서의 객관식 문제는 본초학 과목에서 하위권 학생들의 학업 성취도를 비교적 낮은 난이도의 문항을 사용하여 평가한다는 목적을 충분히 만족하고 있었다.

이처럼 수 개의 문항 형태 또는 평가 내용을 조합할 경우에는 문항들의 타당도 지표가 조화로울 수 있도록 사전에 조율되어야 한다. 특히, 교수법과 학습 목표가 상이한 2-3 과목들을 물리적으로 결합하는 통합 교과의 개발과 시행에 있어서 문항의 타당도가 사전에 적극적으로 검토되어야 하는데, 여러 과목을 하나의 교과로 묶어 시험 문제들을 출제할 경우 구성된 학습 목표(또는 교과)에 따라 선호되는 평가 방법과 난이도의 조합이 뜻하지 않은 부작용을 불러일으킬 수도 있기 때문이다.

예를 들어, 상위권 학생들의 변별력을 중시하는 H 교수와 하위권 학생들에 애정이 많은 L 교수의 통합 교과라면, 고득점을 지향하는 학생들은 학습 목표 H를, 과락 만을 피하려는 학생들은 학습목표 L을 더 중시하게 된다. 이는, 통합 교과에 포함된 학습 목표와 교수자의 조합이 학생들의 학업성취도를 예상치 못한 방향으로 왜곡시킬 수 있다. 또한, 사진과 동영상을 활용하는 멀티미디어 문항과 지식 역량의 측정을 위한 A형, K형 선다형 문항을 함께 사용할 경우, 다양한 임상 현장에서의 한의사 직무역량을 다면적으로 측정하기보다는 특정 유형의 문항에 대한 쏠림과 왜곡된 문항 타당도를 유발할 수도 있을 것이다.

넷째, 문항분석 및 검사이론이 한의학 교육 및 한의사 국가시험에 시급히 도입되어야 하며, CBT 및 CAT의 개발에 적극적으로 활용되어야 한다[7,24,32]. 시험은 교육의 최종 결과가 아니라 학생들의 학습을 안내하는 또 하나의 과정이다. 시험에 있어서 공정한 평가가 더욱 중시되고 있는 현실을 고려한다면, 한의학 교육의 업그레이드를 위한 LMS 및 CBT의 도입 과정에 IRT 분석의 포함이 최우선으로 고려되어야 할 것이다[14,22,24,33].

한의사 국가시험의 문항 개수는 기존의 420개에서 380개를 거쳐 최근에는 340개까지 축소되었다. 필기형 시험에서 CBT로의 형식적 전환, 단순 지식에서 추론과 문제해결로 평가 중심의 이동, 특정 과목의 추가나 삭제 등과 같은 시험의 전체

적인 구조와 방향 및 문항 개수의 수정에는 검사이론을 활용한 검사동등화가 반드시 선행[34]되어야 한다. 명백한 사전 검토와 근거를 마련하지 못하고 진행되는 문항 개수의 획일적 축소나 구조의 임의적이고 급격한 변경은 한의학 교육에 예상치 못한 부작용을 유발할 수 있다.

이에, 본 연구와 같은 분석과정을 통해 시험 문항을 지속적으로 분석, 수정하여 활용한다면, 보다 높은 수준의 한의학 교육이 가능할 것이다. 이와 함께, 문제은행을 공동으로 개발하고 활용하는 과목별 컨소시엄 또는 대학간 협력 시스템을 구축할 수 있다면 교육자원 개발을 위한 시간과 노력의 효율성을 높일 수 있을 것이다[3,5,33]. 본초학은 중요한 기초학 지식 역량의 하나로서 임상 교육과 한의사 직무에서 기술 역량의 기본적인 토대가 된다. 이에, 각 한의과대학에서 재학생들이 일정 수준 이상의 본초학 지식 역량을 지닐 수 있도록 교육하는 것이 중요하며, 교육을 담당하는 교수자 개개인의 교육 경험이나 관심 연구 분야에 상관없이 재학생들이 일정한 본초학 학업성취도를 유지하여야 한다. IRT를 사용하여 전국 재학생들의 본초학 학업 역량을 분석한다면 단원별 부족 또는 충족 여부를 객관적으로 분석하여 보다 표준화된 교육을 전국적인 수준에서 제공할 수 있을 것이며, CTT를 사용할 때 직면할 수 있는 학생 또는 학교를 줄세우기한다는 불필요한 비판도 피할 수 있을 것이다.

본 연구에서는 한의계에 생소한 IRT를 한의과대학에서의 시험분석에 처음으로 활용하였는데, IRT는 최근에 들어서야 한의학 임상도구의 개발 및 타당화에 도입되었다[18-20]. 한의학에 있어서 오랫동안 중시되어온 한의학 고전 문헌에서의 진단 논리를 고려한다면, 이같은 '검사(test) 자체를 평가(analyze)한다'는 개념은 납득하기 쉽지 않다.

검사(진단)와 문항(증상)에 대한 분석은 이제 시작되는 단계로[18-20], 예를 들어, '太陽病, 發熱, 汗出, 惡風, 脈緩者, 名爲中風' (『傷寒雜病論』「太陽病」 제2조)에서는 임상 현장에서의 중풍(中風) 진단이 특정 증상(sign)들의 있음/없음 만을 평가 대상으로 하고 있으며, 확률적 사고와 측정 개념에 대한 통계학적 고려는 포함되어 있지 않다.

그러나, IRT를 토대로 하는 과학적인 임상 검사는 진단을 위해 증상을 측정하는 구체적인 방법, 예를 들어 발열의 측정부위(겨드랑이, 이마, 손목, 구강 등)나 측정 온도(38도, 39도, 40도 등)와 함께, 진단 과정에서 특정 증상의 발현 빈도(예를 들어, 난이도)와 임상적 중요성(예를 들어, 변별도)이 다른 증상들(汗出, 惡風, 脈緩)과 어떻게 차이를 보이는지에 대한 타당화된 수치를 요구한다[20].

이에, 미래 한의사의 기본 역량(competence)인 비판적 사고(critical thinking)에 대한 교육에도 IRT와 같은 새롭고도 분석적인 사고 방식이 적극적으로 도입되어야 한다[8]. 현재 한의학계에서의 통계학 교육은 단순히 ANOVA, χ^2, Correlation과 같은 단순한 분석법과 소프트웨어의 사용법만을 주요 대상으로 사용하고 있는데[28], 본 연구에서 제시된 분석적이며 통계적인 사고는 미래 한의학을 위한 핵심적인 임상 역량이라고 할 수 있다.

다만, 본 연구에서 확인된 내용들을 일반화하기에는 다음과 같은 제한점들이 있으므로 이를 극복하기 위한 후속 연구가 필요하다. 첫째, 본 연구에서 사용된 문항 타당도 지표의 해석은 목적이나 상황에 따라 달라질 수 있으므로, 추가적인 연구를 통해 한의학 교육 분야별로 적절한 해석법이 개발되어야 한다[3,8]. 예를 들어, 보건의료인 국가시험원[31]은 CTT 문항 변별도에서 0.15 이하를 불량, 0.15 ~ 0.25는 경계, 0.25 ~ 0.35는 양호, 0.35 이상을 우수를 사용하였으나, 연구[35]에 따라서 0.2 미만은 변별력이 '거의 없음', 0.21 ~ 0.40은 '있음', 0.40 이상은 '아주 높음'으로 해석하기도 한다.

둘째, 본 연구에서는 문항에 대한 분석만을 다루었으며, 이러한 문항들이 하나로 모아진 시험에 대한 분석을 진행하지는 못하였다. 이에, 시험으로 묶어진 다음에는 어떠한 타당도를 지니고 있는지에 대한 추가적인 연구가 시급히 진행되어야 할 것이다. 이를 통해, 시험 및 문항의 타당도가 성별이나 연령, 학제(학석사 또는 석사 과정), 입학 년도, 교육 과정, 교수자 및 평가자의 특성 등에 의해서 어떻게 변하는지 세밀하게 분석하여야 할 것이다[3,8].

셋째, 본 연구를 통해 기초의학 교육 과정에서의 문항타당도 분석이 유용하다는 것이 확인되었는데, 한의학 임상 교육 과정에서의 다양한 임상 견습, 술기능력 교육, 오수혈 등 임상 지식의 암기 시험, 임상수행평가(CPX)와 객관구조화진료시험(OSCE) 등에서도 유용하다는 것을 재확인하는 후속 연구가 시급하다[3,8,10].

본 연구에서 처음으로 제시된 IRT를 기반으로 한 문항분석은 CBT의 도입을 앞둔 이러닝 시대 근거기반 한의학 교육학의 핵심적인 연구 도구이므로[27,29], 한의학 교수자들이 용이하게 활용할 수 있는 기반이 시급하게 마련되어야 할 것이다[14,22,23].

Ⅴ. 결 론

한의학 교육에 있어서 선다형 시험 문항이 오랫동안 사용되어왔으나, 계산법과 결과 이해의 어려움으로 인하여 최신 IRT가 도입되지 못하여 왔다. 본 연구에서는 기존의 CTT와 최신 IRT를 동시에 사용하여 본초학 시험에 활용된 14문항의 정답 선택 비율, 문항-총점 상관성, 난이도, 변별도, 추측도를 분석한 결과 다음과 같은 결론을 얻었다.

1. 14개 암기형 문항은 4개의 A형 문항과 10개의 K형 문항으로 구성되어있었으며, 정답 선택 비율은 52.8 ~ 97.2%의 범위에 있었다.

2. CTT를 사용한 난이도 분석에서 6개의 문항이 허용범위(0.3 ~ 0.7)에 해당하였고, 변별도 분석에서는 11개의 문항이 양호(0.25 이상)에 해당하였다.

3. IRT를 사용한 난이도 분석에서 4개의 문항이 보통(-0.5 ~ +0.5)에 해당하였고, 변별도 분석에서는 12개의 문항이 적절한(0.65 ~ 1.34) 범위에 해당하였으며, 추측도에 있어서는 7개의 문항이 경계(0.2 ~ 0.3) 수준에 해당하였다.

4. CTT를 사용한 난이도 및 변별도는, IRT를 사용한 난이

도 및 추측도와는 높은 상관성을 보였지만, IRT 변별도와는 유의한 상관성을 보이지 않았다.

이러한 결과는 근거기반 교육을 위한 문항타당도 분석이 CTT 만으로는 충분치 않기에 IRT가 시급히 도입되어야 한다는 것을 의미하며, 본 연구에서의 다면적 문항분석 방법이 시험 문항의 개발, 사용 및 개정의 기본적인 근거로 활용되어야 한다는 것을 시사한다.

감사의 글

이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

References

1. Schauber SK, Hecht M. How sure can we be that a student really failed? On the measurement precision of individual pass-fail decisions from the perspective of Item Response Theory. Medical Teacher. 2020 ; 42(12) : 1374-84.
2. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Medical education. 2010 ; 44(1) : 109-17.
3. Lee SY, Lee Y, Kim MK. Effectiveness of Medical Education Assessment Consortium Clinical Knowledge Mock Examination (2011-2016). Korean Med Educ Rev. 2018 ; 20(1) : 20-31.
4. Lim EY, Park JH, Kwon I, Song GL, Huh S. Comparison of item analysis results of Korean Medical Licensing Examination according to classical test theory and item response theory. J Educ Eval Health Prof. 2004 ; 1(1) : 67-76.
5. Huh S. Application of Computerized Adaptive Testing in medical education. Korean J Med Educ. 2009 ; 21(2) : 97-102.
6. Jeong G-H, Yim MK. Applicability of Item Response Theory to the Korean nurses' licensing examination. J Educ Eval Health Prof. 2005 ; 2(1) : 23-9.
7. Yim MK, Huh S. Testing unidimensionality and goodness-of-fitness for the application of Item Response Theory to the Korean medical licensing examination. Korean J Med Educ. 2007 ; 19(2) : 163-9.
8. Lim H-S, Lee Y-M, Ahn D-S, Lee J-Y, Im H. Item analysis of Clinical Performance Examination using Item Response Theory and Classical Test Theory. Korean J Med Educ. 2007 ; 19(3) : 185-95.
9. Department of test question development in Korea Health Personnel Licensing Examination Institute. Guide for developing test items in Korea Health Personnel Licensing Examination Institute. Seoul, Korea : Korea Health Personnel Licensing Examination Institute. 2013 : 43-51.
10. Park JC, Kim KS. A comparison between discrimination indices and Item-Response Theory Using the Rasch model in a clinical course written examination of a medical school. Korean J Med Educ. 2012 ; 24(1) : 15-21.
11. Park C. Polytomous Item Response Theory Model. Seoul : Educational Science Publishing. 2001 : 11-18.
12. Seong T. Understanding and application of Item Response Theory. Paju : Educational Science Publishing. 2016: 18-53.
13. Meyer JP. Applied measurement with jMetrik. New York : Routledge. 2014 : 40-52, 82-107.
14. Park J-W, Jang L-C, Choi J-W, Lee S-J. The experience of web-based test in medical education. Korean J Med Educ. 2006 ; 18(2) : 183-192.
15. Armstrong R, Belov D, Weissman A. Developing and Assembling the Law School Admission Test. Interfaces. 2005 ; 35(2) : 140-51.
16. Schauber SK, Hecht M, Nouns ZM. Why assessment in medical education needs a solid foundation in modern test theory. Adv in Health Sci Educ. 2018 ; 23(1) : 217-32.
17. Choi Y-J, Asiilkalkan A. R package for Item Response Theory Analysis : description and features. Measurement : Interdisciplinary research and perspectives. 2019 ; 17(3) : 168-75.
18. Lee S, Lee Y, Han SY, Bae N, Hwang M, Lee J, Chae H. Urinary Function of the Sasang Type and Cold-Heat Subgroup Using the Sasang Urination Inventory in Korean Hospital Patients. Evid Based Complement Alternat Med. 2020 ; 2020 : 7313581.
19. Chae H, Cho YI, Lee SJ. The Yin-Yang personality from biopsychological perspective using revised Sasang Personality Questionnaire. Integr Med Res. 2021 ; 10(1) : 100455.
20. Lee Y-j, Lee S, Kim S-h, Lee J, Chae H. Study on the revision and clinical validation of the Sasang Digestive Function Inventory. J Sasang Constit Med. 2021 ; 33(3) : 54-71.
21. Korean Laws Information Center. ACT ON DEVELOPMENT OF E-LEARNING INDUSTRY AND PROMOTION OF UTILIZATION OF E-LEARNING. [cited 2021, Oct 10] Available from https://www.law.

go.kr/lsInfoP.do?lsiSeq=177153&lsId=009747&chrClsCd=010202&urlMode=lsInfoP&viewCls=lsInfoP&efYd=20151210&vSct=ACT%20ON%20DEVELOPMENT%20OF%20E-LEARNING%20INDUSTRY%20AND%20PROMOTION%20OF%20UTILIZATION%20OF%20E-LEARNING&ancYnChk=0#0000

22. Park JH, Son JY, Kim S. Experiences with establishing and implementing Learning Management System and Computer-Based Test System in medical college. Korean J Med Educ. 2012 ; 24(3) : 213-22.
23. Ruiz JG, Mintzer MJ, Leipzig RM. The Impact of E-Learning in Medical Education. Academic Medicine. 2006 ; 81(3) : 207-12.
24. Chae Y-m, Park SG, Park I. The relationship between classical item characteristics and item response time on computer-based testing. Korean J Med Educ. 2019 ; 31(1) : 1-9.
25. Korea Health Personnel Licensing Examination Institute. Computer Based Test. 2021 [cited 2021, Oct 10] Available from https ://www.kuksiwon.or.kr/cnt/c_2033/view.do?seq=11
26. Im E-J, Lee W-K, Lee Y-C, Choe B-H, Chung S-K, Lee T-H, Cho H, Sohn J-H, Won D-I, Kong H-H, Chang B-H, Lee J-M. Development of Computer-Based Test (CBT) and student recognition survey on CBT. Korean J Med Educ. 2008 ; 20(2) : 145-54.
27. Chae H, Hwang S, Kwon Y, Baik Y, Shin S, Yang G, Lee B, Kim JK, Lee B. Study on the Prerequisite Chinese Characters for Education of Traditional Korean Medicine. J Physiol & Pathol Korean Med 2010 ; 24(3) : 373-9.
28. Lee Y, Kwak M-J, Jung H, Ha H-y, Chae H. A study on the statistical methods used in KCI listed journals of traditional Korean medicine from 1999 to 2008. J Korean Orient Med. 2012 ; 18(2) : 55-64.
29. Lee J, Han JH, Kim MS, Lee HS, Han SY, Lee SJ, Chae H. Teaching Yin-Yang biopsychology using the animation, "Pororo the Little Penguin". Eur J Integr Med. 2020 ; 33 : 101037.
30. Han SY, Kim HY, Lim JH, Cheon J, Kwon Y, Kim H, Yang GY, Chae H. The past, present, and future of traditional medicine education in Korea. Integr Med Res. 2016 ; 5(2) : 73-82.
31. Korea Health Personnel Licensing Examination Institute. Development of multiple choice test item. In. Improvement of item development ability (Workshop). Yangsan campus, Pusan National University: Korea Health Personnel Licensing Examination Institute. 2019 : 29-31.
32. Korea Health Personnel Licensing Examination Institute. Item Analysis on the 75th National Licensing Examination (2020) for Korean Medicine Doctor. Seoul: Korea Health Personnel Licensing Examination Institute. 2020 : 2-4, 10-50.
33. Huh S. Computer-based testing and construction of an item bank database for medical education in Korea. Korean Med Educ Rev. 2014 ; 16(1) : 11-15.
34. Huh S. Test equating of a medical school lecture examination based on Item Response Theory : a case study. Korean J Med Educ. 2005 ; 17(1) : 15-28.
35. Kang MJ, Kim MS. Item analysis using Classical Test Theory and Item Response Theory, validity and reliability of the Korean version of a Pressure Ulcer Prevention Knowledge. J Korean Biol Nurs Sci. 2018 ; 20(1) : 11-19.