대한한의학회지 제45권 제1호(2024년 3월)
J Korean Med. 2024;45(1):*1-16*
http://dx.doi.org/10.13048/jkm.24001
pISSN 1010-0695 • eISSN 2288-3339

Original Article

교육 현장에서 시행된 임상 술기 시험의 다면적 타당도 분석

채한[1], 이민정[2,3], 김명호[4], 김규석[5], 조은별[6*]

[1]부산대학교 한의학과, [2]서울대학교 의과대학 의학교육학교실, [3]서울대학교 대학원 휴먼시스템의학과
[4]우석대학교 한방병원 한방내과, [5]경희대학교 한의과대학 한방안이비인후피부과, [6]한국한의학연구원 한의과학연구부

Multifaceted validity analysis of clinical skills test in the educational field setting

Han Chae[1], Min-jung Lee[2,3], Myung-Ho Kim[4], Kyuseok Kim[5], Eunbyul Cho[6]

[1]School of Korean Medicine, Pusan National University
[2]Department of Medical Education, Seoul National University College of Medicine
[3]Department of Human Systems Medicine, Seoul National University
[4]Department of Internal Korean Medicine, Woosuk University Medical Center
[5]Department of Ophthalmology, Otorhinolaryngology, and Dermatology of Korean Medicine, College of Korean Medicine, Kyung Hee University
[6]KM Science Division, Korea Institute of Oriental Medicine

Introduction: The importance of clinical skills training in traditional Korean medicine education is increasingly emphasized. Since the clinical skills tests are high-stakes tests that determine success in national licensing exams, it is essential to develop reliable multifaceted analysis methods for clinical skills tests in actual education settings. In this study, we applied the multifaceted validity evaluation methods to the evaluation results of the cardiopulmonary resuscitation module to confirm the applicability and effectiveness of the methods.
Methods: In this study, we used internal consistency, factor analysis, generalizability theory G-study and D-study, ANOVA, Kendall's tau, descriptive statistics, and other statistical methods to analyze the multidimensional validity of a cardiopulmonary resuscitation test in clinical education settings over the past three years.
Results: The factor analysis and internal consistency analysis showed that the evaluation rubric had an unstable structure and low concordance. The G-study showed that the error of the clinical skills assessment was large due to the evaluator and unexpected errors. The D-study showed that the variance error of the evaluator should be significantly reduced to validate the evaluation. The ANOVA and Kendall's tau confirmed that evaluator heterogeneity was a problem.
Discussion and Conclusion: Clinical skills tests should be continuously evaluated and managed for validity in two steps of pre-production and actual implementation. This study has presented specific methods for analyzing the validity of clinical skills training and testing in actual education settings. This study would contribute to the foundation for competency-based evidence-based education in practical clinical training.

Key Words *: Clinical skills test, CPR examination rubric, multifaceted validity, Generalizability Theory*

・Received : 1 September 2023 ・Revised : 6 December 2023 ・Accepted : 16 February 2024
・Correspondence to : Eunbyul Cho
KM Science Division, Korea Institute of Oriental Medicine
1672, Yuseong-daero, Yuseong-gu, Daejeon, 34054, Republic of Korea
Tel : +82-42-869-2779, Fax : +82-42-861-5800, E-mail : chostar427@gmail.com

서 론

보건의료인의 국가시험은 시험의 결과가 피평가자에게 중요한 영향을 미치는 고부담 시험(high-stakes testing)이다[1]. 그러므로, 명확한 성취목표에 따라 기획되고 안정적으로 진행되어야 하며, 치밀한 역량 평가를 토대로 합격/불합격이 공정하게 결정되어야 하며, 시험 실행 이후에는 종합적인 타당도와 신뢰도 평가를 근거로 지속적인 개선이 이루어져야 한다[2,3].

최근 들어 임상역량 및 술기 수행 기술에 대한 관심이 증가하면서, 기존의 필기시험과 함께 병력청취, 신체진찰, 환자교육, 의사소통 등을 대상으로 하는 임상술기시험(clinical skills test)이 요구되고 있다[4]. 국내 보건의료에 있어서, 의사 국가시험 술기시험은 2009년 아시아에서는 처음으로 도입되었으며[5], 치과의사의 경우 10여년간의 준비 끝에 2021년부터 실시되었다. 한의사에 있어서도 술기시험의 타당성이 논의되고 있다[6].

한의학 교육에 있어서도 진료수행평가와 객관구조화진료시험 등 임상술기시험의 체계적 시행에 대한 연구가 점차 증가하고 있다. 그러나 기존의 연구들이 임상술기시험에 대한 학습자의 반응, 시험 문항의 산발적 개발과 실행 등에 초점을 맞추고 있을 뿐, 시험과 문항 자체의 타당도와 신뢰도에 대한 고민은 매우 부족하다[6,7]. 국가시험에 새로운 평가를 추가하기 위해서는, 평가 목표와 시험에 대한 청사진이 먼저 확립되고, 타당도와 신뢰도를 고려한 시험 문항의 개발이 이루어진 후에, 시뮬레이션을 통해 교육현장에 적용될 수 있는지 확인되어야 한다[5].

이에 한의학 임상술기시험의 개발 및 시행 관리에 필요한 신뢰도와 타당도 분석법이 정립되어야 하며, 교육 현장에서의 구체적인 적용과 활용에 대한 연구가 요구된다. 그러나, 기존의 임상술기시험의 타당도 분석을 살펴보면, 현장의 맥락이나 통계적 가정(assumption)과 무관하게 사용할 수 있는 것으로 오해되어 왔다. 예를 들어, 연속형(예를 들어 1~10점) 평가 지표에 있어서 크론바흐 알파(Cronbach's α)는 단일 차원성과 오차의 독립성을 전제로 검사나 시험의 일관성에 흔하게 사용되었고[8], 범주형(예를 들어 합격/불합격) 평가지표에 있어서는 평가자간 신뢰도(inter-rater reliability) 분석에 Cohen's κ가 무의식적으로 사용되었다. 평가자와 시험장이 한 개로 고정되어 있는 평가 루브릭 개발 과정에서만 적용 가능한 이러한 분석법을 시간과 자원의 제한으로 다수의 학생들을 복수의 평가자와 복수의 평가 스테이션 및 평가 그룹에 임의로 배정하는 실제 임상 교육 현장에 그대로 활용할 수는 없다[9]. 이러한 경우, 연속형 평가지표에서는 순위척도를 사용하는 Kendall's τ를, 그리고 범주형 평가지표에서는 일반화된 Fleiss's κ가 사용되어야 하기 때문이다[10,11].

아울러, 시험의 신뢰도를 단순히 높다/낮다고 분석하는 이러한 사후분석만으로는 평가 현장을 기획, 운영, 개선하기에는 효용성이 매우 부족하기에, 술기시험의 타당도 분석은 구체적인 평가 루브릭, 평가자, 스테이션 등의 다양한 현장 상황이 유발할 수 있는 타당도 저하를 정량적으로 평가하면서 개선 방법까지 제시하여야 한다[12-14]. 타당도의 일반화(Generalizability, G) 과정에서 특정 국면과 국면의 상호작용이 만들어내는 오차 또는 타당도 저하의 크기를 계산하는 G-study와 시뮬레이션을 통해 오차의 크기를 줄임으로써 일반화된 타당도 향상법을 결정(Decision, D)하는 D-study로 구성된 일반화가능도 이론(Generalizability Theory)은 이 같은 요인(국면, facet)의 영향력을 정량적으로 분석하는 데 적합하다[12,13].

이와 함께, 평가 루브릭이 기획과정에서 의도한 임상역량 구조를 유지하는지 확인할 요인 분석, 평가자 및 스테이션 등에 따른 평가 점수의 차이는 없는지를 확인할 분산분석[9], 평가 그룹별로 점수의 분포에 차이가 없는지를 비교할 왜도와 첨도, 분산 및 최소점수 같은 기술통계[12], 그리고 총점 및 타 항목과의 관련성을 검토하기위한 상관분석[12] 등이 요구된다.

국가시험은 피평가자가 특정 영역에서의 지식과

술기 역량이 최저 기준인 준거(criterion)를 충족하는지를 확인하는 준거참조평가(criterion-referenced evaluation)이기에, 준거로서의 합격점수 설정이 시험의 공정성에 직결된다[15]. 타당성을 담보하기위한 합격/불합격 점수를 설정하는 방법으로는 고전검사이론을 근거로 한 modified Angoff 방법, 문항반응이론을 근거로 한 Bookmark 방법 등이 제시되고 있으나[16], 국가시험 또는 교육 현장에서는 편의를 위하여 60점(100점 만점)을 관습적으로 사용하여 왔다. 만약, 실제 교육현장에서의 평가 결과를 토대로 공정한 준거 점수를 제안한다면, 평균 합격률을 고려한 %ile 점수와 합격-불합격 그룹간의 유의한 차이를 고려한 99%CI 등이 유용하게 활용될 수 있을 것이다.

이상에서 언급한 내적일치도, 일반화 가능도, 분산분석 등을 포함한 다면적 타당도 분석법은 한의학 교육현장에서 실제 시행되었던 결과를 대상으로 적용해볼 때, 교육의 실질적인 개선을 이끌어 낼 수 있을 것이다. 아울러, 현장에서의 시행 결과에 대한 객관적 평가는 환류 과정을 통한 교육 프로그램 개선의 핵심적인 요소이므로, 기존에 한의학 교육에 활용되고 있는 임상술기시험에 대한 적용과 검토가 필요하다.

이에, 본 연구에서는 일개 한의학과에서 십여년 이상 운영되어 온 임상술기시험 중에서 매년 지속적으로 시행되어 온 심폐소생술(CPR) 모듈의 3년간 평가 결과에 위에서 제시된 다면적 타당도 평가방법을 적용해 봄으로써 활용 가능성과 유효성을 확인해 보고자 하였다. 학습자 역량 분석 측면에서 평가의 타당성을 높이기 위한 노력이 바로 역량중심, 근거기반 술기교육을 지속하는 데 필요조건인 만큼, 본 연구에서 제시된 다면적 평가 타당도 분석 방법은 한의학과 및 보건의료인국가시험원에서 임상 술기 평가를 도입할 때 기여할 수 있을 것이다.

연구 대상 및 방법

1. 연구 대상

본 연구에서는 한의학과에서 3년(2019-2021) 동안 시행된 2일간의 술기 시험에서 CPR 시행 결과만을 연구 분석에 활용하였다. 매 시행 연도마다 학생들을 임의의 그룹(8개)으로 나누어 임의의 평가자(4명)에게 배정하였으며, 연도별로 동일(16문항) 또는 일부를 삭제한(14문항) CPR 평가 루브릭을 사용하였다. CPR 평가의 타당도 분석에는 연도별로 사용된 평가 문항별 채점 점수를 사용하였으며, 평가 총점의 특성 분석에는 연도별 채점 총점(17점 또는 20점)을 100점으로 환산하여 사용하였다. 본 연구는 생명윤리위원회의 심의(PNU IRB/2022_75_HR) 이후 한의학교육실에서 데이터를 받아 진행하였다.

2. 연구 방법

1) 평가 점수에 대한 기술통계 분석

100점 만점으로 환산된 평가 총점에 대한 기술통계 결과를 시행 연도, 평가 그룹 및 평가 그룹별로 제시하였으며, 평가 그룹별 평균, 표준편차, 중앙값, 분산, 왜도, 첨도, 최소값, 최대값을 제시하였다. 점수 분포의 비대칭성을 측정하는 왜도(Skewness)가 -1보다 작은 경우 평가 점수가 고득점으로 상당히 치우쳐 있음을 의미하며, 점수 분포의 꼬리 길이와 중앙부위의 뾰족함을 측정하는 첨도(Kurtosis)의 값은 3보다 클 경우 정규분포보다 뾰족한 것 또는 특정 점수로 편향된 것으로 해석하였다. 분산은 평가 점수의 분포가 넓게 퍼진 정도로서 분산이 작을 경우 특정 점수로 편향되어 있는 것으로 해석하였으며, 최소점은 가장 낮게 평가된 평가 점수로 고득점으로의 편향을 분석함에 활용하였다.

2) 내적일치도 분석

평가 루브릭을 구성하는 문항들이 동일한 평가 목표를 지니고 있는지 확인하기 위한 내적일치도 분석

에는 3개 시행 연도별로 Cronbach's α를 시행하였으며, 평가 문항 중 응답치의 분산이 0인 경우 분석에서 제외하였다. 평가 루브릭에 대한 분석에 있어서는 채점 총점의 평균, 표준편차와 분석에 사용된 문항 및 피험자의 숫자, 그리고 Cronbach's α를 제시하였으며, α값은 0.7이상을 유의미한 것으로 해석하였다. 평가 문항에 대한 분석에 있어서는 채점 점수의 평균, 분산, 문항-잔여문항의 합 사이의 상관성, 문항 삭제시 α값을 제시하였으며, 분산이 비교적 크거나 문항-잔여문항의 합과의 상관성이 정적(positive)이며 높거나 문항 삭제시 α값이 낮아지는 문항을 유의미한 것으로 해석하였다.

3) 요인 분석

평가 루브릭을 구성하는 문항들이 안정적인 구조를 지니고 있는지 확인하기 위하여 탐색적 요인분석을 시행하였으며, 최소잔차법을 사용하여 추출된 요인들을 대상으로 Varimax 회전을 시행하였다. 요인의 개수를 선정함에 있어서는 Eigenvalue가 1 이상인 것을 기준으로 하였으며, 필요한 경우 Eigenvalue의 스크리 도표에서 급격한 변화를 보이는 지점을 활용하였다. 요인 모형의 적합도 지표로는 Root Mean Square Error of Approximation(RMSEA), Tucker-Lewis Index(TLI), Bayesian information criterion(BIC) 그리고 χ^2를 사용하였으며, RMSEA가 0.1보다 작거나 TLI가 0.9보다 크거나 χ^2 분석결과 유의할 경우 유의미한 모델로 해석하였다. 이와 함께 문항별 요인부하량이 0.3 이상일 경우 평가 문항이 해당 요인에 포함되는 것으로 제시하였으며, 0.3 이하의 요인부하량은 생략하였다.

4) 일반화가능도 이론

일반화가능도 이론(Generalizability theory)는 관측된 점수의 변동에 대한 개별 원인 및 상호작용의 영향력을 분석하여 관측이나 검사를 일반화할 수 있는지를 분석한다.

일반화 연구(Generalizability study, G-study)는 술기시험에서의 평가자, 피험자, 평가 문항 등과 같은 다양한 요인(국면, facet) 및 이들의 상호작용이 평가 결과에 미치는 영향을 분석하는 데 사용된다. 본 연구에서는 학생(participant, p)과 평가자(rater, r)의 교차모형을 사용하였으며, 각 국면 및 상호작용의 제곱평균(sum of square), 제곱평균(mean of square) 및 분산성분(variance component)을 구하였다.

국면의 추정된 분산성분은 평가 점수를 전집점수로 일반화하는 데 따르는 오차의 크기 또는 영향력으로 해석되는데, 국면 또는 잔차의 상대적인 영향력 크기는 분산성분 값을 서로 비교하여 결정한다. 분산성분의 값이 음수(-)로 나오는 것은 표집 과정에서의 오차가 과도하게 크거나 측정 모형을 제대로 명시하지 못하였기 때문이며[12], 본 연구에서는 이를 0으로 대체하여 처리하여 결정 연구에 활용하는 Cronbach의 접근 방법을 사용하였다.

결정 연구(Decision study, D-study)는 안정성 있는 평가 결과를 얻기 위한 술기시험 시행모델의 구체적인 개선방법을 제시하기 위한 것으로, G-study에 사용된 오차 국면의 차원(또는 크기)을 증가시켜 오차분산을 축소시킴으로써 일반화가능도 계수(generalizability coefficient)를 임의로 증가시켜 볼 수 있다.

앞에서 G-study를 통해 얻은 전집분산($\sigma^2(\tau)$)을 기준으로 상대 결정을 위한 상대오차 분산($\sigma^2(\delta)$)과 절대결정을 위한 절대오차 분산($\sigma^2(\Delta)$)을 변화시키는 과정에서, 상대결정을 위한 일반화가능도 계수(generalizability coefficient, $E\rho^2$)와 절대결정을 위한 의존도 계수(dependability coefficient, Φ)를 도출하였다. 국가시험과 같은 합격-불합격을 판단하기 위한 준거 참조 평가(criterion-referenced evaluation)에서 타당화된 평가 결과의 의존도 계수(Φ) 값은 0.7 이상이 되어야 하며, 본 연구에서는 이를 만족할 때까지 평가자 국면의 차원을 임의로 증가시켰다.

5) CPR 평가 총점의 차이에 대한 분석

시행 연도(2019, 2020, 2021년) 및 평가자(총 9명)에 따른 CPR 총점의 유의한 차이를 분산분석(ANOVA)을 사용하여 분석하였으며, Levene's test 결과에 따라 Bonferroni 또는 Holm을 사용한 사후검정을 시행하였다. 이와 함께, 시행 연도와 평가자에 따른 CPR 평가 총점의 분포 차이를 직관적으로 확인할 수 있도록 히스토그램과 밀도함수로 제시하였는데, 두 그룹의 분포 형태의 유사성이 높을수록 차이가 없는 것으로 해석하였다.

특정 국면 또는 오차요인에 의한 점수분포의 차이가 추정될 경우에는 평가 총점을 순위척도로 활용하는 Kendall tau-b를 사용하여 평가자간 일치도 분석을 시행하였다. Kendall의 τ 계수는 -1~+1범위를 보이며, 0.8 이상일 경우에 평가자간 신뢰도가 충족된 것으로 해석하였다[13].

6) CPR 평가에서의 합격점수 산정

타당성과 신뢰성을 지닌 합격점수를 설정하기 위하여, 3년간 시행된 CPR 평가 총점에 대한 기술통계 분석을 시행하였다. 3년동안의 CPR 평가 총점의 중앙값, 평균, 표준편차, 표준 오차, 왜도, 첨도, 그리고 하위 10%ile, 20%ile 그리고 30%ile 점수, 95%CI 하한값, 99%CI 하한값 점수를 제시하였다.

3. Statistical Analysis

본 연구에서 시행된 일반화가능도 분석에는 urGENOVA 2.1(Robert L. Brennan, The University of Iowa, 2001)을 사용하였으며, 그 외의 모든 분석에는 jamovi 2.3.26(The jamovi project, https://www.jamovi.org)을 사용하였다[17]. 통계 분석의 결과에 있어서 통계적 유의수준으로는 $p<0.05$, $p<0.01$ 및 $p<0.001$을 사용하였으며, 통계 분석 결과치는 소수점 셋째자리까지 제시하였다.

결 과

1. 기술통계 분석

100점 만점으로 환산된 평가 총점에 대한 연도 및 평가 그룹별 기술통계 결과는 Table 1과 같다. 고득점에 치우쳐 있음을 의미할 수 있는 -1보다 작은 왜도는 2019년의 4, 5, 6번 평가그룹에서, 2020년의 3, 5, 6, 8번 평가그룹에서, 2021년의 2, 3, 5, 7번 그룹에서 확인되었다. 특정 점수로의 편향성을 의미할 수 있는 3보다 큰 첨도는 2019년의 6번 평가그룹, 2020년도의 5, 6, 8번 평가그룹, 2021년의 2, 3, 7번 평가 그룹에서 확인되었다.

평가 그룹에 따른 분산의 차이가 큰 것 또는 최저점수에 확연한 차이를 볼 수 있었는데, 2019년의 4, 6, 8번 그룹의 분산은 60 이하로 타 그룹(136.667~274.167)에 비하여 확연히 작았다. 2021년에는 3개의 그룹을 확인할 수 있었는데, 2, 3번 그룹의 분산은 5.802이었고, 1, 5번 그룹은 각각 41.302, 50.428이었으며, 4, 6, 7번은 103.548~116.135인 것으로 확인되었다. 아울러, 낮은 분산을 지닌 평가 그룹에서는 높은 최저점수를, 높은 분산을 지닌 평가 그룹에서는 낮은 최저점수를 지니고 있음을 확인할 수 있었다.

2. 내적일치도 분석

실행 연도별 CPR 평가 루브릭에 대하여 내적일치도를 분석한 결과는 Table 1, 2와 같다. 루브릭 전체에 대한 분석 결과(Table 1), 2019년은 16문항에서 Cronbach's α값이 0.646, 2020년은 13문항에서 0.457, 2021년은 12문항에서 0.545인 것으로 확인되었다. 이처럼 0.7보다 작은 내적일치도 분석 결과(Table 1)는 평가 루브릭을 구성하는 문항들이 동일한 평가 목표를 지니고 있지 못함을 의미한다.

루브릭을 구성하는 문항별 분석 결과(Table 2), 2020년에서는 1번, 5번, 12번 문항에서, 2021년에서는 5번, 13번 문항에서 분산이 0이었기에 분석과정

에서 제외되었다. 2019년에는 6번, 12번 문항이 삭제될 경우 타당도(α값)가 높아지는 것으로, 2020년에는 6번, 13번, 15번 문항이 삭제될 경우 타당도가 높아지며, 2021년에는 1번, 4번, 10번, 12번, 14번 문항이 삭제될 경우 타당도가 높아지는 것으로 확인되었다.

16번 문항은, 2019년에는 높은 분산과 높은 문항-잔여문항합의 상관성, 그리고 삭제시 α값이 감소(0.646에서 0.539로)하였으며, 2020년도에도 삭제시 α값이 감소(0.457에서 0.382로)하였기에 2021년에도 지속적으로 사용되었어야 하는 좋은 평가문항이라 해석할 수 있다(Table 2). 이와 반대로, 6번 문항은, 문항-잔여문항합의 상관성이 2019년과 2020년에서 음수(-)로 확인되었으며, 삭제시 α값이 증가(0.646에서 0.655로, 0.457에서 0484로) 하였기에, 2021년에서는 삭제되었어야 하는 좋지 않은 평가 문항이라고 해석할 수 있다. 이와 함께, 1번 및 5번 문항은 연도별 평균 점수가 0.976, 0.979 또는 1.0으로 변별도가 매우 낮은 문항이므로, 평가 루브릭에서는 생략해도 될 것이다.

3. 요인 분석

실행 연도별 CPR 평가 루브릭에 대한 요인분석을 시행한 결과는 Table 3, 4와 같다.

Table 1. Description and scale reliability statistics of CPR examination in year 2019, 2020 and 2021

Year	Group	n	items	Cronbach's α	Mean	Std.Dev.	Median	Variance	skewness	kurtosis	Min.	Max.
2019		47	16	0.646	17.362	2.566	90.000	164.593	-0.615	-0.607	55	100
	1	6	16		81.667	11.690	82.500	136.667	0.245	0.959	65	100
	2	6	16		75.833	16.558	77.500	274.167	0.128	-0.665	55	100
	3	6	16		87.500	13.693	90.000	187.500	-0.876	-0.048	65	100
	4	6	16		97.500	4.183	100.000	17.500	-1.537	1.429	90	100
	5	6	16		90.833	12.007	95.000	144.167	-1.201	0.847	70	100
	6	6	16		96.667	6.055	100.000	36.667	-1.952	3.657	85	100
	7	6	16		81.667	12.910	77.500	166.667	0.705	-1.623	70	100
	8	5	16		82.000	7.583	80.000	57.500	0.315	-3.081	75	90
2020		49	16	0.457	15.667	1.589	95.000	63.223	-1.104	0.188	70	100
	1	6	16		95.000	6.325	97.500	40.000	-0.889	-0.781	85	100
	2	6	16		90.833	8.612	90.000	74.167	-0.026	-2.367	80	100
	3	6	16		88.333	10.328	90.000	106.667	-1.172	1.970	70	100
	4	6	16		90.000	7.746	90.000	60.000	0.000	-1.875	80	100
	5	6	16		96.667	6.055	100.000	36.667	-1.952	3.657	85	100
	6	6	16		97.500	6.124	100.000	37.500	-2.449	6.000	85	100
	7	6	16		92.500	8.216	95.000	67.500	-0.811	-1.029	80	100
	8	7	16		94.286	8.864	95.000	78.571	-2.215	5.299	75	100
2021		41	14	0.545	13.500	1.569	94.100	83.106	-0.973	0.542	65	100
	1	6	14		90.200	7.101	91.150	50.428	0.086	-1.541	82	100
	2	6	14		99.017	2.409	100.000	5.802	-2.449	6.000	94	100
	3	6	14		99.017	2.409	100.000	5.802	-2.449	6.000	94	100
	4	6	14		91.183	10.343	94.100	106.970	-0.492	-1.928	77	100
	5	6	14		88.217	6.427	88.200	41.302	-1.358	2.467	77	94
	6	6	14		87.267	10.777	88.250	116.135	-0.517	-0.596	71	100
	7	5	14		82.340	10.176	88.200	103.548	-1.932	3.701	65	88

요인에 해당되지 않는 문항 또는 분산이 0인 문항이 다수 존재함을 확인할 수 있었고, 2-3개의 요인에 해당되는 문항들이 수개(1번, 2번, 3번)의 문항을 제외하고 년도별로 상이한 것을 볼 때 요인 구조가 명확치 않았던 것으로 해석할 수 있었다(Table 3). 이와 함께, RMSEA가 0.1보다 컸으며 TLI가 0.9보다 작았기에 평가 루브릭이 불안정한 구조를 지니고 있는 것으로 해석할 수 있었다.

2019년에서는 3개의 요인이 추출되었다(Table 4). 첫번째 요인에는 1번, 2번, 3번, 13번, 14번 문항, 두번째 요인에는 5번, 7번, 8번, 9번, 10번, 16번 문항, 세번째 요인에는 4번, 6번, 9번, 12번, 13번 문항이 해당되었으며, 10번 및 15번 문항은 세가지 요인 모두에 대한 요인부하량이 0.3보다 작으면서 고유량이 0.9 이상이었다.

2020년에서는 2개(5번, 12번) 문항에서 분산이 0이었기에 제외되었으며, 2개의 요인이 추출되었다. 첫번째 요인에는 2번, 3번, 16번 문항이 해당되었으며, 두번째 요인에는 9번, 10번, 14번 문항이 해당되었으며, 1번, 4번, 6번, 7번, 8번, 11번, 13번, 15번 문항은 두가지 요인 모두에 대한 요인 부하량이 0.3보다 작으면서 고유량이 0.9 이상이었다. 스크리 도

Table 2. Item reliability statistics of CPR examination rubric in year 2019, 2020 and 2021

item	2019				2020				2021			
	Mean	SD	Item-rest correlation	α if deleted	Mean	SD	Item-rest Correlation	α if deleted	Mean	SD	Item-rest correlation	α if deleted
1	0.979	0.146	0.140	0.644	1.000	0.000	-	-	0.976	0.156	0.056	0.549
2	0.979	0.146	0.140	0.644	0.980	0.143	0.063	0.456	0.951	0.218	0.322	0.509
3	0.979	0.146	0.140	0.644	0.980	0.143	0.063	0.456	0.927	0.264	0.381	0.490
4	0.745	0.441	0.190	0.641	0.857	0.354	0.142	0.443	0.976	0.156	0.056	0.549
5	0.979	0.146	0.319	0.635	1.000	0.000	-	-	1.000	0.000	-	-
6	0.979	0.146	-0.094	0.655	0.959	0.200	-0.106	0.484	0.927	0.264	0.243	0.519
7	0.809	0.449	0.368	0.613	0.837	0.373	0.133	0.447	0.780	0.419	0.382	0.471
8	1.681	0.515	0.507	0.585	1.898	0.306	0.172	0.435	1.707	0.461	0.499	0.422
9	0.830	0.433	0.298	0.625	0.918	0.277	0.359	0.390	0.902	0.300	0.317	0.500
10	1.213	0.907	0.414	0.621	1.673	0.658	0.356	0.351	0.976	0.156	0.056	0.549
11	0.957	0.204	0.161	0.642	0.959	0.200	0.233	0.429	1.683	0.471	0.308	0.498
12	1.000	0.209	0.000	0.653	1.000	0.000	-	-	0.927	0.264	0.047	0.557
13	1.809	0.449	0.260	0.630	1.898	0.421	-0.007	0.498	1.000	0.000	-	-
14	0.957	0.204	0.118	0.645	0.939	0.242	0.359	0.398	1.780	0.419	-0.056	0.608
15	0.936	0.247	0.257	0.634	0.939	0.242	0.064	0.459				
16	1.532	0.654	0.656	0.539	1.816	0.391	0.323	0.382				

Table 3. Model fit measures of factor analysis using CPR rubric items in year 2019, 2020 and 2021

Year	RMSEA	TLI	BIC	χ^2	df	p-value
2019	0.583	-0.047	987.208	1275.969	75	<.001
2020	0.487	-0.055	559.066	808.143	64	<.001
2021	0.136	0.358	-83.077	76.606	43	0.001

RMSEA, Root Mean Square Error of Approximation; Tucker-Lewis Index, TLI; Bayesian information criterion, BIC.

표를 기준으로는 한 개의 요인을 지니는 것으로 확인되었으며, 이 경우 RMSEA는 0.509, TLI는 -0.135, BIC는 757.753, χ^2는 1057.424($p<0.001$)이었다.

2021년에서는 2개(5번, 13번) 문항에서 분산이 0이었기에 제외되었으며, 2개의 요인이 추출되었다. 첫번째 요인에는 2번, 3번, 9번, 12번 문항이 해당되었으며, 두번째 요인에는 6번, 7번, 8번, 11번 문항이 해당되었고, 1번, 4번, 10번, 14번 문항은 두가지 요인 모두에 대한 요인 부하량이 0.3보다 작으면서 고유량이 0.9 이상이었다.

4. 일반화가능도 이론(Generalizability Theory)

본 연구에서 사용된 학생(p)과 평가자(r)의 교차모형에 대한 일반화 연구(G-study)와 결정 연구(D-study)의 분석 결과는 Table 5, 6과 같다. G-study 분석 결과(Table 5)에 있어서 2020년의 결과만 학생(p) 국면의 분산성분이 양(+)의 값을 보였는데, 이는 다른 시행연도(2019년 및 2021년)와는 다르게 표집 과정에서의 오차가 허용할 수 있는 범위 안에 있으며, 2020년도 결과에 대해서만 D-study를 추가적으로 진행할 수 있음을 의미한다. 2020년도 잔차(학생과 평가자의 상호작용)의 분산성분 값(660.928)이 학생 국면에 비하여 매우 큰 것을 확인할 수 있는데, 이는 표현되지 않은 요인에 의한 영향력이 크다는 것으로 해석할 수 있다.

안정성 있는 CPR 술기시험의 수정방법에 대한 D-study는 2020년도에서만 시행되었으며(Table 6), 평가자(r) 국면의 차원을 증가시키거나 양질의 평가자를 통해 오차분산을 줄임으로써 얻을 수 있는 일반화가능도 계수($E\rho^2$)와 의존도 계수(Φ)를 확인하였다. 의료인 면허 국가시험과 같은 고부담 시험에서의 타당화된(의존도 계수 0.7 이상) 결과를 위해서는 평가자의 차원을 현재의 4명에서 22명으로 5배수 이상 증가해야 했는데, 이는 평가자로 인한 오차를 크게 줄이거나 평가자의 균질성을 매우 높여야 한다는 것

Table 4. Factor loading of CPR rubric items in year 2019, 2020 and 2021

	2019				2020			2021		
	Factor			Uniqueness	Factor		Uniqueness	Factor		Uniqueness
	1	2	3		1	2		1	2	
1	0.994			0.002			0.990			0.997
2	0.994			0.002	0.997		0.005	0.935		0.121
3	0.994			0.002	0.997		0.005	0.821		0.309
4			0.451	0.725			0.985			0.946
5		0.426		0.815	-	-	-	-	-	-
6			0.927	0.131			0.995		0.308	0.892
7		0.546		0.649			0.997		0.926	0.134
8		0.714		0.419			0.990		0.542	0.669
9		0.348	-0.464	0.663		0.967	0.065	0.364		0.839
10		0.503		0.740		0.725	0.467			0.952
11				0.951			0.997		0.460	0.779
12			-0.599	0.627	-	-	-	0.386		0.839
13	0.633		0.342	0.458			0.992	-	-	-
14	0.688			0.507		0.602	0.637			0.995
15				0.950			0.918			
16		0.917		0.159	0.304		0.893			

 http://dx.doi.org/10.13048/jkm.24001

을 의미한다.

5. CPR 평가 총점에 영향을 미치는 요인

앞에서 실시한 일반화가능도 분석에서 제시한 국면(facet)별 오차를 초래한 특성이 무엇인지 확인하기 위하여, 술기시험의 시행 연도 및 평가자에 따른 CPR 평가 총점의 차이를 분석하였다.

시행연도에 따른 CPR 총점의 유의한 차이를 확인한 결과, F(2,134) 값은 4.857로 유의하였으며(p=0.009), 사후분석 결과 2019년과 2020년의 총점 사이에 유의한(p=0.008) 차이를 확인할 수 있었다. 평가자에 따른 CPR 총점의 유의한 차이를 확인할 결과, F(8, 128) 값은 6.411로 유의하였으며(p<0.001), 3번 평가자가 5번, 6번, 8번, 9번 평가자와, 4번 평가자가 9번 평가자와, 8번 평가자가 9번 평가자와 유의한 차이를 보였다.

시행 연도 및 평가자에 따른 CPR 총점의 분포를 히스토그램과 밀도함수로 제시하여 분산분석에서 확인된 연도 및 평가자에 따른 차이가 CPR 평가점수 프로파일에서 기인한다는 것을 확인할 수 있었다 (Figure 1). 특정 연도(2020년) 및 평가자(9번 평가자 등)에서 평가 점수 분포가 고득점으로 치우침을 확인할 수 있었는데, 이러한 결과는 평가자 및 시행 연도 국면이 일반화가능도 이론에서의 큰 오차분산을 유발한 원인이었음을 확인하는 것이다.

평가자에 의한 유의한 차이가 확인되었으므로 켄

Table 5. Analysis results of G-study in year 2019, 2020 and 2021

Year	Effect	degree of freedom	T	Sum of Square	Mean of Square	Variance Component
2019	p	40	2843703.645	2214042.579	55351.064	-3489.952
	r	3	651972.027	22310.961	7436.987	-1509.119
	pr	120	11183319.360	8317304.754	69310.873	69310.873
	Total	163		10553658.294		
2020	p	48	60633.713	45420.887	946.268	71.335
	r	3	15293.153	80.327	26.776	-12.942
	pr	144	155887.719	95173.679	660.928	660.928
	Total	195		140674.893		
2021	p	46	189166927.415	152602872.108	3317453.741	-179236.126
	R	3	37879212.951	1315157.643	438385.881	-76510.901
	Pr	138	747229042.693	556746957.635	4034398.244	4034398.244
	Total	187		710664987.385		

Table 6. Analysis results of D-study in year 2020

Size of facets		Variance components			Coefficients	
student	rater	$\sigma^2(\tau)$	$\sigma^2(\delta)$	$\sigma^2(\Delta)$	$E\rho^2$	Φ
1	4	71.335	165.232	165.232	0.30	0.30
1	8	71.335	82.616	82.616	0.46	0.46
1	12	71.335	55.077	55.077	0.56	0.56
1	21	71.335	31.473	31.473	0.69	0.69
1	22	71.335	30.042	30.042	0.70	0.70
1	23	71.335	28.736	28.736	0.71	0.71

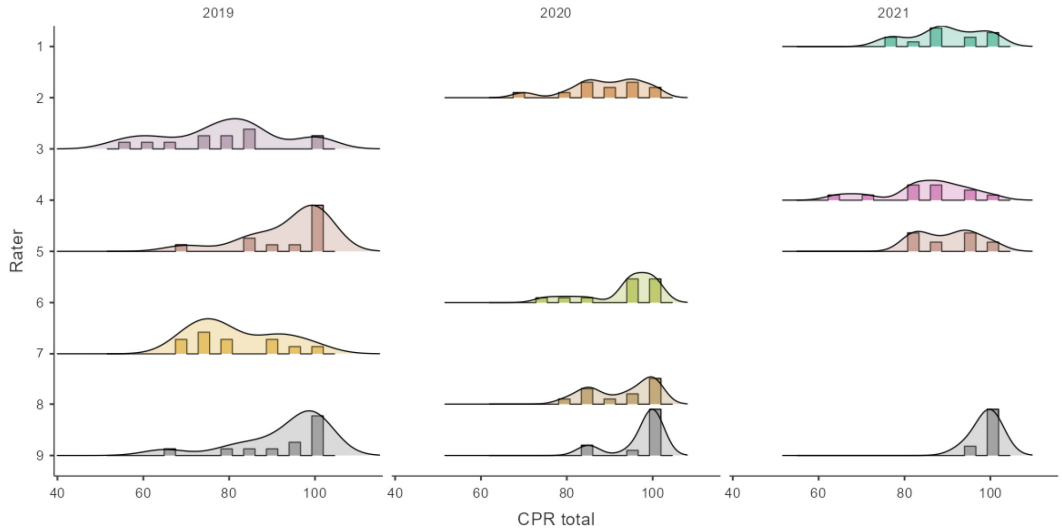

Fig. 1. CPR total score according to the year and rater.

달의 tau-b 분석을 통해 평가자 신뢰도를 확인한 결과, τ 계수는 2019년에는 0.190(t=1.605, p=0.109), 2020년에는 0.305(t=2.538, p=0.011), 2021년에는 0.398(t=3.042, p=0.002) 이었으며, 3개년도 모두에 있어서는 0.274(t=4.179, p<0.001)로 확인되었다. 이러한 결과는 2020년, 2021년 그리고 3개년 종합에서 τ 계수가 유의하였으나 모두 기준이 되는 0.8보다 매우 낮았다.

6. CPR 평가에서의 합격-불합격 점수 설정

고부담 시험으로 활용되는 CPR 술기시험의 합격-불합격 점수를 타당하게 설정하기 위하여, 3년동안의 CPR 평가 점수에 대한 기술통계를 시행하였다.

137명의 3년간 CPR 시험 점수에 있어서, 중앙값은 94.100, 평균은 90.409, 표준편차는 10.481, 표준오차는 0.895, 왜도는 -1.026, 첨도는 0.460으로 확인되었다. 이와 함께, 하위 10%ile는 75.0점, 20%ile는 82.4점 그리고 30%ile는 85.0점으로 확인되었으며, 95%CI의 하한값은 88.639점, 99%CI 하한값은 88.07점으로 확인되었다.

국가시험에서의 합격률이 90% 이상인 것을 고려한다면 합격-불합격 구분 점수로 75점이 적절할 것이며, 평가 참여자들과 통계적으로 유의한 차이(99%CI)를 보이는 평가 점수를 구분 점수로 활용한다면 88점이 적절할 것으로 확인되었다.

고찰 및 결론

본 연구는 일개 한의과대학에서 3년간 시행한 CPR 임상 술기시험 결과를 분석하고, 한의학 임상술기시험의 신뢰도와 타당도 평가방법을 구체적으로 제시하였다.

기술통계 분석 결과 평가 그룹별 분산의 차이가 컸으며, 일부 평가그룹의 점수의 고득점 편향이 원인이었다. 평가 루브릭 전체의 내적일치도는 3개년 시험 모두 Cronbach's α가 모두 0.7 미만으로, 의료분야 선행 임상술기시험에 비해 낮은 편이었으며[18,19], 그 원인을 문항별 분산, 문항-잔여문항합의 상관성, 문항 삭제시 α값을 활용하여 확인하였고, 평가 루브릭이 불안정한 구조도 요인분석을 통해 확인하였다.

일반화가능도 이론을 활용하여 평가점수에 영향을 미치는 요인들을 분석한 결과, G-study를 통해 평가자 및 상호작용의 분산성분과 표현되지 않은 요인의 영향력이 매우 크며[20], D-study를 통해 평가자 오차를 크게 줄여야 할 필요성을 확인하였다. 특히 켄달의 τ 계수로 확인한 평가자 신뢰도가 0.8보다 매우 낮아, 평가자 신뢰도를 높일 방안이 시급하다는 것을 확인할 수 있었다.

한의학교육에 있어서 임상술기시험의 평가에는 주로 전문가 안면타당도[14]가 사용되었으며, 체계적으로 신뢰도와 타당도를 분석한 선행연구는 찾아보기 어렵다.[6,7] 본 연구에서는 교육 현장에서 실제 시행되었던 CPR 술기시험의 다면적 타당도를 요인분석, 내적일치도, 일반화 가능도 이론, 분산분석 등을 활용하여 검토하였으며, 임상술기 평가를 위한 시험과 문항의 개발과 운영에 시급한 개선이 필요하다는 것을 통계적으로 확인할 수 있었다. 이와 함께 많은 비용과 인력이 필요한 임상술기시험의 본격적인 도입에 앞서 실제 교육현장에서 선제적으로 준비해야 할 타당도 분석법과 예상되는 문제점들의 검토를 미리 준비할 수 있었다.

일개 한의대에서 시행된 CPR 임상 술기시험의 결과를 토대로 전국적인 국가시험 수준에서 임상 술기를 만들고 시행하기에 앞서 다음과 같은 것들이 고려되어야 한다.

첫째, 임상 술기 시험의 시행 목적을 명확히 해야 한다. 시험 목적 설정은 OSCE 개발 시 가장 첫 단계에서 이루어져야 하는데, 본 연구의 임상술기시험과 같은 저부담시험은 고부담시험에 비해 높은 신뢰도가 가장 중요한 것으로 고려되지 않으며 응시자에게 피드백을 제공하는 목적이 크기 때문이다. 본 연구의 대상은 형성평가로서의 의미가 강하기에 낮은 타당도와 평가루브릭의 불안정한 구조가 문제를 초래할 가능성이 비교적 적을 수도 있겠지만, 국가고시와 같은 고부담 시험에서는 높은 공정성을 담보할 수 있는 방법이 마련되어야 한다.

둘째, 사후 분석을 통해 평가 루브릭이 잘 만들어져야 한다. 본 연구에서 시험결과 및 문항에 대한 기술통계(Table 1), 내적일치도 분석(Table 1, 2), 요인분석(Table 3, 4)을 통해 사후 분석의 중요성을 확인할 수 있었다. 평가 그룹에 따라 점수의 왜도, 첨도가 균질하지 않은 점은 평가의 공정성 측면에서 평가 그룹의 구성을 통제할 필요가 있음을 시사한다. 또한 사후 분석을 통해 내적일치도 분석 상 타당도가 낮은 문항은 삭제하거나, 실제 임상 술기에서 중요한 항목이 요인분석 상 요인 부하량이 높은 문항이 되도록 평가 루브릭을 재구성할 필요가 있다.

셋째, 임상 술기시험의 공정하고 타당한 진행을 위한 연구와 준비가 필요하다. 현재까지 한의학 교육에 있어서 임상 술기시험 연구는 한의사로서 임상현장에서 요구되는 역량의 설계와 이를 신뢰롭게 평가할 수 있는 평가 루브릭의 개발에 초점을 두고 있었다. 반면, 학생들을 대상으로 임상 교육 현장에서 학생들을 대상으로 임상술기 교육과 평가를 시행함에 있어서는 평가자간 채점 결과의 차이까지 고려하여 공정하고 타당하게 관리하는 것이 매우 중요하다[12,21].

평가 루브릭의 개발 과정에서는 기존에 사용되고 있는 타당도 지표를 그대로 사용할 수 있다. 예를 들어, 연속형 평가 지표(예, 0~10점)를 위한 Cronbach's α와 불연속 평가 지표(예, 합격/불합격)를 위한 Cohen's κ는 평가 루브릭 개발 과정에서 소수의 가상 피평가자와 평가자에 적용함에 특별한 문제가 되지는 않을 것이다. 다만, 본격적인 교육현장 적용에 앞선 시범사업 및 대학별 평가에 있어서는 다수의 평가자와 복수의 스테이션이 활용될 것이므로, 연속형 평가지표에 있어서는 Cronbach's α 대신 Kendall's τ가 사용되어야 하며, 불연속 평가지표에 있어서는 Cohen's κ 대신 Fleiss's κ가 활용되어야 할 것이다. 이와 함께, 임상 술기 시험 과정에서의 오차 원인과 그 영향력의 크기를 G-study와 D-study를 사용하여 분석함으로써 문제점을 파악하고 개선 방안을

도출하여야 하며, 특정 요인의 영향력에 대한 상세한 분석을 위해 상관분석, 분산분석 및 기술통계 등이 사용되어야 할 것이다.

넷째, 공정한 합격점수를 어떻게 설정할 것인가에 대해 논의하여야 한다. 본 연구에서는 3년간 총점을 활용하여 학생 점수의 순위(10%ile) 또는 분포(99%CI)를 기준으로 합격점 산정방법을 제시하였다. 합격점(cut score)이란 절대평가로 알려진 준거참조평가(criterion-referenced evaluation)에서 준거를 실무적인 점수로 나타낸 것이다. 준거참조평가에서는 학습자가 특정 영역에서 얼만큼 지식과 술기 수준에 도달하였는지 알아보기 위하여 최저기준인 준거(criterion)를 설정하여 평가한다[22]. 공정하고 신뢰할 수 있는 시험 결과를 위해서는 준거와 합격점이 타당하게 설정되는 것이 중요하다.

타당성 있는 준거설정 방법에는 대표적으로 고전검사 이론(Classical test theory; CTT)을 근거로 한 modified Angoff 방법, 문항 반응 이론(IRT)를 근거로 한 Bookmark 방법, 학생의 수행을 중심으로 평가하는 경계선 회귀(Borderline regression) 방법 등이 있다[23]. 임상술기 시험에서 주로 활용되는 경계선 회귀 방법에서는 평가 체크리스트의 총점과 전반적인 등급 점수 사이에 회귀분석으로 합격, 불합격을 결정한다[24,25]. 필기시험에서 가장 널리 사용되는 modified Angoff 방법에서는 패널을 구성하여 합격점을 평정하고 실제 시험 결과, 합격률 데이터를 참고하여 수렴, 합의하는 과정을 반복하여 진행한다[16]. 국가시험과 일개 대학의 임상술기 시험에서 타당한 준거설정을 지속하기 위해서는 다양한 준거설정 방법을 모의 시행, 비교해보고, 타당하면서 지속할 수 있는 방법을 고안할 필요가 있다.

다섯째, 평가결과의 유의한 해석을 위해서 타당도 확보는 필수이므로[26], 임상 술기 시험 또한 이러한 타당도 확보를 위해 시행 결과를 매년 평가하고 평가 결과를 바탕으로 시험을 개선해야 한다. 만약 타당도가 다른 문제들을 임의로 배정하거나 차등화된 학점을 제시할 경우, 평가 결과 적용에 대한 공정성이 문제될 수 있다. 따라서, 이러한 타당도 확보를 위해 본 연구에서 제시한 다면적 타당도 분석 방법을 고려할 수 있다. 또한, 임상 술기 시험 관련 문항 개발 및 평가를 위해 전문 학회에서 검토한 필수역량 측정의 타당성과 신뢰성을 확보한 통일된 기준을 한의계가 사용한다면 운영과 관리가 용이할 것이다. 한의학 교육 전반에 대한 체계적인 연구 및 학술교류를 목적으로 2023년 발족한 한의학교육학회가 주도적으로 역할을 할 수 있을 것이다. 이와 같은 술기 시험의 평가 및 개선을 포함한 전반적인 한의학 교육 과정의 개발, 운영, 평가, 개선 등에 주도적 역할을 담당할 수 있는 각 대학별 교육실의 설립 및 활동이 절실하다.

본 연구의 한계와, 제안하고자 하는 후속연구는 다음과 같다.

첫째, CPR 1개 과목만을 대상으로 하였는데, 다른 스테이션의 경우 모든 학생들이 응시하지 않아 일관된 분석이 어려웠기 때문이다. 그러나 OSCE는 일반적으로 여러 스테이션으로 구성되어 학생들이 순차적으로 모든 스테이션을 경험하는 시험이다. 전형적인 OSCE는 20개 이상 스테이션으로 구성되며[27], 기존 연구에서 스테이션 개수는 4개부터 40개까지 다양하였다[28]. 본 연구에서는 단일 과목 시험 결과만을 분석하였으나, 3년간의 데이터를 활용하여 평가 루브릭을 구성하는 문항들의 내적일치도, 일반화가능도, 평가자간 일치도를 확인할 수 있었다. 향후 여러 과목으로 구성된 OSCE를 실행한 후 시험을 평가하여 목적 적합성, 교육에 미친 영향, 개선할 점을 도출해야 한다[27].

둘째, CPR 과목은 매년 지속적으로 실시되었으며 시험이 이틀간 진행되어 학생들이 시험에 응시하기 이전에 시험정보가 노출되었을 가능성이 있다. 다만, 본 연구에서는 학생의 학습량 차이보다 평가자 등에 의한 차이가 더 컸고, 가장 낮은 점수(55점)와 왜도(skewness)가 -1보다 작지 않은 그룹도 존재하여 평

가점수가 고득점으로 치우치지 않는 경우들을 확인하였다. 따라서 시험정보가 사전에 주어졌더라도, 술기를 연습하지 않는 학생들도 있다고 추측할 수 있다. 동일한 OSCE 스테이션을 늦게 응시하는 학생은 시험에 통과할 가능성이 더 높으며, 그 중에서도 능력이 부족해서 원래 시험에 통과시키지 말았어야 할 학생은 특히 더 유리하다[29]. OSCE의 공정성을 위해서는 모든 학생들을 격리시키고, 하루 내에 모든 시험이 완료되도록 가능하면 여러 장소에서 동시에 시험을 진행하는 방안을 고려할 수 있다[30]. 향후 한의학교육에서 OSCE의 공정성을 담보하면서도 실행가능한 방안에 대한 체계적인 연구가 필요하다.

셋째, OSCE 시행 과정에서 타당도를 높이기 위한 개선방법에 대한 연구가 필요하다. 의사 실기시험과 관련하여 의과대학 컨소시엄이 구축되어 표준화환자 사례 개발 및 훈련을 공동으로 진행하는 것처럼, 문항 개발은 많은 전문가가 참여해야 하는 어려운 작업이다. 임상 술기 시험의 문항 개발, 표준화환자 훈련, 시험 실행, 평가의 질 관리에 요구되는 수많은 비용도 중요한 문제로 지적되고 있다. 술기시험의 모범답안을 예상하며 시험문제와 평가기준을 개발하는 것은 비교적 쉬운 일로 여겨질 수 있으나, 개발한 문항을 교육현장에서 실행할 때 예측하지 못한 채점표 오류, 환자 교육 부족, 응시자의 돌발행동 등 다양한 문제들에 직면할 수 있다. 임상 술기 시험 평가목표 및 문항, 필수적인 사전 교육 시간, 교육 내용, 스테이션의 개수, 평가자 및 환자 훈련 방법 등 시험 개발에 대한 종합적인 연구와, 시험을 실행한 후 신뢰도와 타당도를 분석하여 시험을 개선하는 연구 모두 필요하다.

넷째, 술기시험 문항 개발-실행-사후 분석의 방법론을 확립해야 한다. 현재 한의학 술기 시험 연구는 산발적으로 각 대학내에서 개별 교수자 주도로 문항이 개발되고 있으며, 문항 개발 방법도 상이하다[7]. 바쁜 교육현장에서 다양한 술기 시험을 효율적으로 실행하기 위해서는 컨소시엄을 구성하여, 확립된 프로세스에 따라 문항을 공동으로 개발한 후, 각 대학에서 실행해야 한다. 선행연구[1]에서 모의고사의 타당도 분석법을 제시하였고, 본 연구에서는 술기 시험의 타당도를 분석하였다. 연구결과를 종합적으로 활용하여, 역량중심 한의학교육에 활용할 수 있는 방안을 검토해야 할 것이다. 한의학교육의 전문 학회 중심으로 문항 개발부터 시험에 대한 사후 분석까지 포괄하는 가이드라인을 만들어 한의과대학(원)에서 동일한 문항을 실행하면, 다량의 데이터를 확보하여 더욱 체계적인 분석이 가능할 것이다.

본 연구에서는 실제 교육 현장에서 3년간 시행된 술기시험 결과의 타당도를 분석하였다. 체계적이고 다면적인 분석법이 제시되고 활용되었으며, 개선이 필요한 부분을 명확히 확인할 수 있었다. 이는 술기 교육에서의 역량중심 & 근거기반 교육의 토대 마련에 기여할 것이며, 평가와 개선과 같은 교육실의 기본 역할 정립에 도움이 될 것이다.

Conflict of interest

We authors declare no potential conflict of interest relevant to this article.

Acknowledgement

본 연구는 부산대학교의 연구비지원을 받았음.

This work was supported by a 2-Year Research Grant of Pusan National University.

Ethical statement

This study was approved b the Institutional Review Board (Consent No.: PNU IRB/2022_75_HR).

Data availability

The data of this study are available upon reasonable request.

참고문헌

1. Chae H, Cho E, Kim SK, et al. Analysis on validity and academic competency of mock test for Korean Medicine National Licensing Examination using Item Response Theory. *Keimyung Medical Journal* 2023. DOI: .
2. Park SH. Possibilities and Limits of High Stakes Testing in US. *Korean Journal of Comparative Education* 2010; 20: 1-21.
3. Eggen TJ and Stobart G. High-stakes testing–value, fairness and consequences. *High-Stakes Testing in Education*. Routledge, 2015, pp.1-6.
4. Korea Health Personnel Licensing Examination Institute. Clinical skill test, https://www.kuksiwon.or.kr/EngHome/cnt/c_3109/view.do?seq=18 (2023, accessed 2023-06-18 2023).
5. Kim KS. Introduction and administration of the clinical skill test of the medical licensing examination, republic of Korea (2009). *Journal of Educational Evaluation for Health Professions* 2010; 7.
6. Han SY, Lee S-H and Chae H. Developing a best practice framework for clinical competency education in the traditional East-Asian medicine curriculum. *BMC Med Educ* 2022; 22: 352. 2022/05/11. DOI: https://doi.org/10.1186/s12909-022-03398-4.
7. Shin J, Go Y, Song C, et al. Presentation on research trends and suggestion for further research and education on Objective Structured Clinical Examination and Clinical Performance Examination in Korean Medicine education: Scoping review. *Society of Preventive Korean Medicine* 2022; 26: 87-112. DOI: 10.25153/SPKOM.2022.26.2.008.
8. Korean Laws Information Center ACT ON DEVELOPMENT OF E-LEARNING INDUSTRY AND PROMOTION OF UTILIZATION OF E-LEARNING. In: Ministration of Trade, Industry and Engergy, (ed.). *18358*. Sejong, Korea: Korean Laws Information Center, 2021.
9. Chae H, Han SY, Yang G, et al. Study on the herbology test items in Korean medicine education using Item Response Theory. *Kor J Herbology* 2022; 37: 13-21. DOI: https://doi.org/10.6116/kjh.2022.37.2.13.
10. Chae H, Lee SJ, Han c-h, et al. Study on the Academic Competency Assessment of Herbology Test using Rasch Model. *J Korean Med* 2022; 43: 27-41. DOI: https://doi.org/10.13048/jkm.22017.
11. Kang Y. Evaluating the cutoff score of the advanced practice nurse certification examination in Korea. *Nurse Education in Practice* 2022; 63: 103407. DOI: https://doi.org/10.1016/j.nepr.2022.103407.
12. Kim S and Kim Y. *Generalizability Theory*. 2nd ed. Paju: Education Science Publishing, 2016.
13. Nunnally JC and Bernstein IH. *Psychometric theory*. 3 rd ed. New York: McGraw-Hill, 1994.
14. Shin S, Kim GS, Song JA, et al. Development of examination objectives based on nursing competency for the Korean Nursing Licensing Examination: a validity study. *J Educ Eval Health Prof* 2022; 19: 19. 2022/08/23. DOI: https://doi.org/10.3352/jeehp.2022.19.19.

15. Seong T, Kang DJ, Kang E, et al. *Introduction to Modern Pedagogy*. Seoul: Hakjisa, 2018.
16. Lee M-j. Exploring the feasibility of implementing criterion-referenced assessment in Korean medicine education: enhancing comprehension and relevance. *J Kor Med Edu* 2023; 1: 10-14. DOI: https://doi.org/10.23215/JKME.PUB.1.1.10.
17. Chae H. Jamovi, an open-source software for teaching data literacy and performing medical research. *J Kor Med edu* 2023; 1: 28-36. DOI: https://doi.org/10.23215/JKME.PUB.1.2.28.
18. Navas-Ferrer C, Urcola-Pardo F, Subirón-Valera AB, et al. Validity and reliability of objective structured clinical evaluation in nursing. *Clinical Simulation in Nursing* 2017; 13: 531-543.
19. Hur HK, Park SM, Kim KK, et al. Evaluation of Lasater judgment rubric to measure nursing student' performance of emergency management simulation of hypoglycemia. *Journal of Korean Critical Care Nursing* 2012; 5: 15-27.
20. Kim J and Cho L-R. Analysis of error source in subjective evaluation on patient dentist interaction: Application of Generalizability Theory. *The Journal of the Korean Dental Association* 2019; 57: 448-455.
21. Lee SY, Lm SJ, Yune SJ, et al. Assessment of Medical Students in Clinical Clerkships. *Korean Medical Education Review* 2013; 15: 120-124.
22. Rim MK, Ahn D-S, Hwang IH, et al. *Validation study to establish a cutoff for the national health personnel licensing examination*. 2014. Korea Health Personnel Licensing Examination Institute.
23. Ahn S and Choi S. Proposal for a Cut Score for the Physics Ability Test: Comparison between the Modified Angoff, Bookmark, and IDM Methods. *New Physics: Sae Mulli* 2018; 68: 599-610. DOI: http://dx.doi.org/10.3938/NPSM.68.599.
24. Schoonheim-Klein M, Muijtjens A, Habets L, et al. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education* 2009; 13: 162-171. DOI: https://doi.org/10.1111/j.1600-0579.2008.00568.x.
25. Pell G, Fuller R, Homer M, et al. How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. *Medical Teacher* 2010; 32: 802-811. DOI: 10.3109/0142159X.2010.507716.
26. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003; 37: 830-837. DOI: 10.1046/j.1365-2923.2003.01594.x.
27. Harden RM, Lilley P and Patricio M. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment*. Elsevier Health Sciences, 2015.
28. Patrício MF, Julião M, Fareleira F, et al. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical teacher* 2013; 35: 503-514.
29. Ghouri A, Boachie C, McDowall S, et al. Gaining an advantage by sitting an OSCE after your peers: a retrospective study. *Medical Teacher* 2018; 40: 1136-1142.
30. Iyer A and Dovedi V. Is there such a thing as

a fair OSCE? *Medical Teacher* 2018; 40: 1192-1192.

ORCID

채한 https://orcid.org/0000-0002-8698-8229
이민정 https://orcid.org/0000-0001-6372-2201
김명호 https://orcid.org/0000-0003-2320-1633
김규석 https://orcid.org/0000-0002-3802-8717
조은별 https://orcid.org/0000-0003-3431-1109