해부 · 생물인류학 제 39 권 제 3 호
Anat Biol Anthropol Vol. 39, No. 3 (2026) pp. 285~300
https://doi.org/10.11637/aba.2026.39.3.285

Check for updates

Original Article

보건의료계열 학생 선발에서 윤리적 판단과 인성·도덕성 관련 수행의 평가: 다면인적성면접에 대한 체계적 문헌고찰

안준무[1], 이승희[1], 김선경[2]

[1]서울대학교 의과대학 의학교육학교실, [2]동의대학교 한의과대학 의학교육학교실

Assessing Ethical Judgment and Character- and Morality-Related Performance in Health Professions Admissions: A Systematic Review of the Multiple Mini-Interview

Junmoo Ahn[1], Seung-Hee Lee[1], Seon-Kyoung Kim[2]

[1]*Department of Medical Education, Seoul National University College of Medicine*
[2]*Department of Medical Education, College of Korean Medicine, Dong-Eui University*

Abstract : Anatomy education provides an early educational context in which health professions students encounter human dignity, respect for life, and professional responsibility through the donated human body. This raises a preceding question for admissions: how can ethical judgment and character- and morality-related performance be observed before students enter educational settings in which respect for human dignity must be enacted in practice? This systematic review examined the validity of the multiple mini-interview (MMI) as a structured approach to observing such performance, rather than judging character as a fixed attribute. Six databases were searched on 9 July 2026, identifying 952 records; 46 full texts were assessed and 45 publications were included. Of these, 33 primary empirical studies constituted the principal evidence base, 11 systematic reviews contextualized the synthesis, and 1 station-construction guide informed design considerations. MMI stations clustered into four ethically salient categories—ethical dilemmas, interpersonal integrity, empathic communication, and public responsibility—interpretable in relation to autonomy, justice, honesty, responsibility, compassion, and the common good. Multi-station, multi-rater MMIs generally showed greater reliability than traditional interviews, and performance was associated with selected proximal outcomes, including clinical and licensing-examination performance. However, evidence that admission-stage ethical or moral performance predicts sustained ethical behavior in clinical practice remained insufficient. Online and generative-AI applications raised additional issues regarding fairness across applicant groups, construct clarity, response authenticity, and human review and accountability. Thus, the MMI is better understood as a structured assessment that samples

저자(들)는 '의학논문 출판윤리 가이드라인'을 준수합니다.
저자(들)는 이 연구와 관련하여 이해관계가 없음을 밝힙니다.
Received: August 31, 2026; **Revised:** September 23, 2026; **Accepted:** September 29, 2026
Correspondence to: 김선경 (동의대학교 한의과대학 의학교육학교실) **E-mail:** pietas@deu.ac.kr

ⓒ 2026 Korean Association of Physical Anthropologists
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
ISSN 2671-566X (Online)

ethical judgment and related professional performance than as a measure of completed moral character. As an interpretive implication, admissions assessment may be aligned with subsequent education in human dignity, medical ethics, and professionalism, including formative encounters with the donated human body in anatomy education. The review was registered in PROSPERO (CRD420261446189).

Keywords : Multiple mini-interview, School admission criteria, Ethics, Professionalism, Medical education

서 론

의학을 배우는 일은 인체구조를 배우는 일에서 시작하고, 그 첫 배움의 중요한 자리는 해부학 실습실이다. 학생들은 그곳에서 한 생을 살아낸 뒤 자신의 몸을 배움의 자리에 내어준 사람을 마주한다. 눈 앞의 몸은 탐구의 대상이지만, 동시에 한 사람의 삶과 죽음을 지나온 몸이다. 해부학 및 해부 실습에 관한 연구들은 이러한 경험이 인체구조에 관한 지식의 습득에 그치지 않고, 기증자와 그 가족에 대한 감사와 존중, 생명과 죽음에 대한 성찰, 책임과 전문직업성에 대한 인식으로 이어질 수 있음을 보여주었다[1-3]. 따라서 해부학 교육은 인체구조를 배우는 자리이면서 동시에 인간의 몸을 어떻게 대해야 하는지를 구체적으로 묻는 교육의 자리이기도 하다.

그러나 바로 이 지점에서 선발과 교육 사이의 긴장이 드러난다. 인간존엄에 대한 존중과 전문직으로서의 책임이 교육과 경험을 거치며 형성되어 가는 것이라면, 그 교육에 들어오기 전 우리는 무엇을 볼 수 있으며 또 어디까지 판단할 수 있는가. 인격 형성은 개별 덕목의 명칭을 알고 그 목록을 늘리는 데 있기보다, 구체적인 관계와 상황 속에서 무엇이 옳고 가능한지를 분별하여 행위로 옮기는 실천적 지혜(phronesis)를 기르는 과정으로 논의되어 왔다[4]. 사람됨을 형성해 가는 교육의 요청과, 아직 형성의 과정에 있는 사람 가운데 누군가를 선발해야 하는 제도적 요구는 바로 이 지점에서 긴장을 이룬다.

따라서 보건의료계열 학생 선발에서 핵심적인 질문은 누가 '좋은 인격'을 갖추었는가를 판정하는 데 있지 않다. 오히려 가치가 충돌하고 타인의 자율성·취약성·존엄을 고려해야 하는 구체적인 상황에서 지원자가 어떠한 판단과 행동을 보이는지, 그리고 그 순간의 수행으로부터 우리는 어디까지 말할 수 있는가가 중요하다. 본 연구에서 '인성·도덕성 관련 수행의 평가'는 인격이나 도덕성을 고정된 속성으로 측정한다는 의미가 아니라, 특정 상황에서 드러나는 윤리적 판단, 의사소통, 타인에 대한 태도와 관련 행동을 제한적으로 관찰하고 해석한다는 의미로 사용한다. 이러한 구분은 이후 다면인적성면접의 타당도와 한계를 해석하는 본 연구의 기본 전제가 된다.

이러한 구분은 보건의료 전문직의 선발에서 특히 중요하다. 의료인은 타인의 생명과 건강뿐 아니라 자율성, 취약성, 존엄과 직접 마주하며, 의료적 판단과 행위에는 기술적 적절성과 함께 윤리적 정당성이 요구된다. 자율성 존중, 선행, 악행금지, 정의는 생명의료윤리에서 판단과 행위를 검토하는 주요 원칙으로 제시되어 왔으며[5], 덕 윤리의 관점에서는 좋은 의료실천이 원칙을 적용하는 데 그치지 않고 구체적인 상황에서 환자의 선을 분별하고 책임 있게 행위하는 실천적 지혜와 관련된다고 본다[6]. 따라서 보건의료계열 선발에서 중요한 것은 지원자가 특정한 덕목을 '소유하고 있는가'를 판정하는 것이 아니라, 가치가 충돌하는 상황에서 문제를 어떻게 인식하고 판단하는지, 타인의 관점과 존엄을 어떻게 고려하는지, 그리고 그 판단을 어떠한 방식으로 의사소통과 행동에 드러내는지를 관찰하는 데 있다.

이러한 수행을 구조화된 방식으로 관찰하기 위해 보건의료계열 학생 선발에서 활용되어 온 방법 가운데 하나가 다면인적성면접(Multiple Mini-Interview, MMI)이다. Eva 등은 지원자의 수행이 상황에 따라 달라질 수 있다는 맥락 특이성(context specificity)을 고려하여, 한 지원자를 여러 독립된 상황과 평가자에 걸쳐 반복적으로 관찰하는 방식의 MMI를 제안하였다[7]. 서로 다른 상황에서 수행을 반복적으로 표집하는 이러한 구조는 한 번의 면접에서 관찰된 수행에 지나치게 의존하는 한계를 줄이기 위한 것이다. 국내에서도 의학전문대학원 학생 선발에 MMI가 도입되었으며, 84명의 지원자를 대상으로 3개의 8분 스테이션을 운영한 연구에서 일반화 가능도 계수 0.791과 지원자 및 면접관의 긍정적인 수용성이 보고되었다[8].

그러나 이러한 다중 표집 구조가 곧 지원자의 인격이나 도덕성 전체를 측정할 수 있음을 의미하지는 않는다. MMI가 직접 제공하는 것은 특정한 상황에서 나타난 판단과 행동의 표본이며, 그 표본으로부터 다른 상황에서의 수행이나 향후 행동을 어느 범위까지 추론할 수 있는가는 별도의 타당도 문제이다. 선행 체계적 문헌고찰에서는 MMI의 신뢰도와 수용성에 대체로 긍정적인 근거가 보고되었으며, 일부 후속 수행

www.kci.go.kr

과의 관련성도 확인되었다[9]. 그러나 이러한 근거는 주로 객관구조화임상시험과 면허시험을 비롯한 교육·평가 성과에 집중되어 있고, 장기적인 예측타당도에 관한 근거는 상대적으로 제한적이다[10]. 특히 입학 단계에서 관찰된 수행이 실제 임상현장에서 지속적인 윤리적 행동으로 이어지는지를 보여주는 근거는 아직 충분하지 않다[9,10]. 따라서 MMI의 타당성을 논의할 때에는 그것이 무엇을 관찰할 수 있는가뿐 아니라, 그 관찰 결과로부터 어디까지 일반화할 수 있는가를 함께 검토해야 한다.

기존 연구는 MMI가 어떠한 비학업적 영역을 평가하는지와 그 신뢰도·타당도·수용성을 폭넓게 검토해 왔다. 체계적 문헌고찰에서는 의사소통, 팀워크·협력, 윤리적·도덕적 판단을 포함한 다양한 개인적 영역이 MMI의 평가 대상으로 사용되고 있음이 확인되었으며[11], 의학교육 선발 전반에 관한 체계적 문헌고찰에서도 MMI는 주요 선발 방법 가운데 하나로 다루어졌으나, 졸업 이후의 수행이나 장기적인 전문직 행동을 뒷받침하는 종단적 근거는 제한적인 것으로 지적되었다[12]. 따라서 이제 필요한 것은 MMI가 유용한 선발도구인가를 다시 묻는 데 그치지 않고, 그 안에서 관찰되는 윤리적 판단과 인성·도덕성 관련 수행이 무엇이며, 그러한 수행을 어떠한 범위에서 해석할 수 있는지를 함께 검토하는 일이다.

이러한 해석의 문제는 면접 환경의 변화와 함께 더욱 중요해지고 있다. 비대면 면접의 확산과 생성형 인공지능의 도입은 MMI의 실행 방식뿐 아니라 무엇을 평가하고 그 결과를 어떻게 신뢰할 것인가라는 문제를 새롭게 제기한다. 최근 실제 의과대학 선발에서 생성형 인공지능으로 개발한 MMI 스테이션이 전통적으로 개발된 스테이션과 유사한 수준의 심리측정적 성능을 보였지만, 연구자들은 공정성과 타당성을 확보하기 위한 지속적인 질 관리의 필요성을 강조하였다[13]. 또한 생성형 인공지능이 지원자의 응답 생성에 개입할 수 있는 선발 환경에서는 평가하려는 구인을 명확히 하고, 평가의 공정성·진정성·책임성을 면접 설계 차원에서 다시 검토할 필요가 있다는 논의가 제기되고 있다[14]. 이러한 변화는 결국 평가도구의 정교함만큼이나 '무엇을 관찰하려는가'와 '관찰된 수행으로부터 어디까지 판단할 것인가'가 중요함을 다시 드러낸다.

이에 본 연구는 보건의료계열 학생 선발에 활용되는 MMI에 관한 문헌을 체계적으로 고찰하여, MMI에서 관찰되는 윤리적 판단과 인성·도덕성 관련 수행의 내용을 분석하고, 그러한 수행의 관찰도구로서 MMI의 신뢰도와 타당도 및 그 해석의 한계를 검토하고자 하였다. 그리고 비대면 MMI와 생성형 인공지능 기반 문항 개발에서 새롭게 제기되는 윤리적·방법론적 쟁점을 함께 살펴보았다. 이를 통해 MMI를 사람의 인격을 판정하는 도구가 아니라 특정 상황에서 나타난 수행을 제한적으로 관찰하는 평가로 이해하고, 선발 단계의 평가와 이후 인간존엄·의료윤리·전문직업성 교육 사이의 관계를 교육적 함의의 수준에서 논의하고자 하였다. 연구 질문은 다음과 같다.

RQ1. MMI 스테이션에서는 어떠한 윤리적 판단과 인성·도덕성 관련 수행이 관찰되며, 이들은 의료윤리의 원칙 및 관련 전문직 가치와 어떠한 관계에서 해석될 수 있는가?

RQ2. 이러한 수행의 관찰도구로서 MMI의 신뢰도와 타당도에 관한 근거는 어떠하며, 특히 입학 단계에서 관찰된 수행을 이후의 교육성과와 장기적인 임상윤리 행동으로 일반화하는 데에는 어떠한 한계가 있는가?

RQ3. 비대면 MMI와 생성형 인공지능 기반 문항 개발은 공정성, 구인 명료화, 응답의 진정성, 인간의 검토와 책임이라는 측면에서 어떠한 윤리적·방법론적 쟁점을 제기하는가?

연구 방법

1. 연구 설계 및 보고 준거

방법과 보고는 의학교육 체계적 문헌고찰의 실무 지침인 AMEE Guide 94 [15]와 STORIES 보고 표준[16]을 준용하고, 문헌 선별은 PRISMA 2020 [17], 검색 전략은 PRISMA-S 확장 지침[18]에 따라 보고하였다. 포함 연구의 교육맥락·스테이션 유형·결과변수가 이질적이어서 효과크기 통합(메타분석)은 수행하지 않았으며, 대신 양적 결과는 SWiM 지침[19]에 따른 방향성 투표로, 질적 범주화는 GRADE-CERQual [20]로 확실성을 평정하였다. 프로토콜 확정 이후의 변경은 '프로토콜로부터의 이탈: 등록 프로토콜과 달리 중복 제거와 제목 · 초록 선별을 단일 통합 단계로 수행하였다.' 본 고찰의 프로토콜은 국제 전향적 체계적 문헌고찰 등록부(PROSPERO)에 공개 등록되었다(등록번호 CRD420261446189).

2. 핵심 구인의 조작적 정의

윤리적 판단과 인성·도덕성 관련 수행은 문헌마다 서로 다른 용어와 의미로 사용되어 왔다. 이 고찰은 자료를 일관되게 추출하고 해석하기 위해 핵심 구인을 다음과 같이 조작적으로 정의하였다(Table 1). 이러한 정의는 인격이나 도덕성을 고정된 속성으로 측정하기 위한 것이 아니라, MMI 스테이션에서 실제로 관찰할 수 있는 판단·행동·언어를 분석하기 위한 기준이다. 따라서 이 연구의 해석은 관찰 가능한 수행의 범위 안에서 이루어지며, 이를 인격이나 도덕성 전체에 대한 판정으로 확대하지 않는다.

www.kci.go.kr

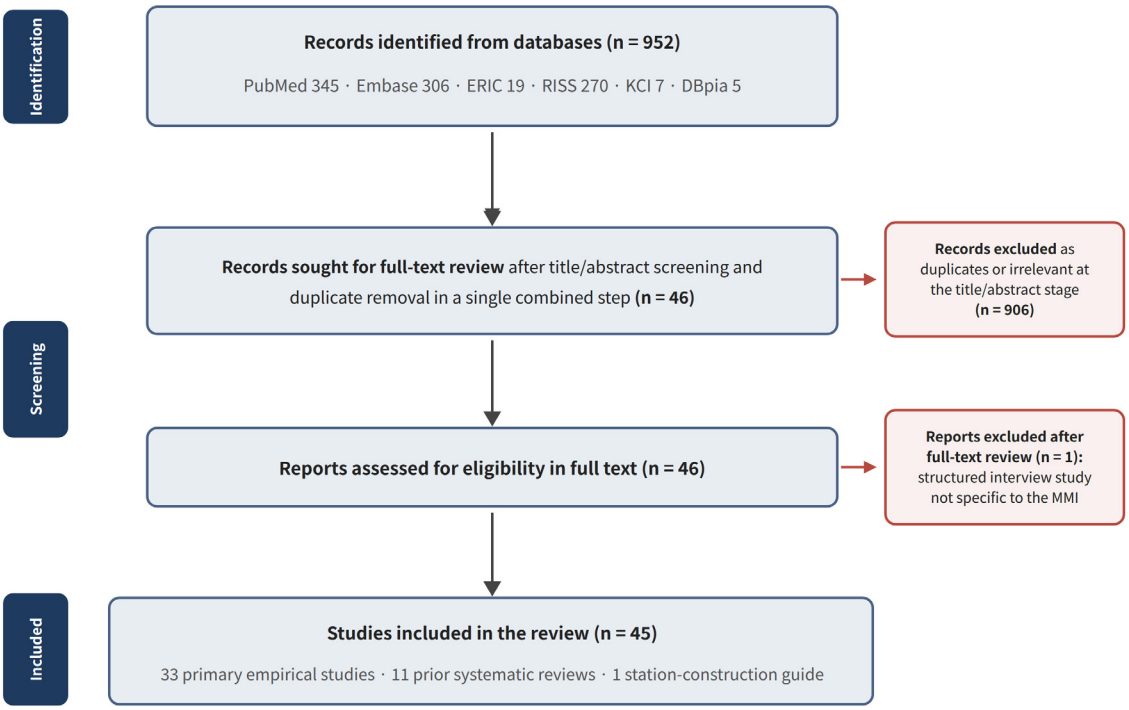


Fig. 1. Records were identified from six databases, screened, and assessed for eligibility, and 45 publications were included: 33 primary empirical studies, 11 systematic reviews, and 1 station-construction guide. Duplicate removal and title/abstract screening were performed as a single combined step, in which 906 duplicate or irrelevant records were excluded. Counts at the identification stage include duplicates across databases. The search was conducted on 9 July 2026. ERIC, Education Resources Information Center; KCI, Korea Citation Index; MMI, multiple mini-interview; RISS, Research Information Sharing Service.

3. 검색원과 검색 전략

국외 데이터베이스는 PubMed·Embase·ERIC을, 국내 데이터베이스는 KCI·RISS·DBpia를 검색원으로 설정하였고 검색일은 모두 2026년 7월 9일이다. 검색어는 MMI 도메인, 선발·입학 도메인, 인성·도덕성 도메인, 신뢰도·타당도 도메인을 조합하였다. 국외 검색식은 ("multiple mini interview" OR "multiple mini-interview" OR MMI)와 (ethics OR character OR moral OR professionalism OR empathy OR virtue OR integrity OR "non-cognitive") 및 (validity OR reliability OR selection OR admission)을 결합하였다(검색식 개발 단계의 예비 검색에서 광의 320건·협의 153건이 확인되어 최종 검색식은 광의 전략을 채택). 국문은(다면인적성면접 OR MMI)와 (인성 OR 도덕성 OR 윤리 OR 공감 OR 인격 OR 의료인문학 OR 전문직업성) 및 (타당도 OR 신뢰도 OR 선발 OR 평가)를 조합하였다. 데이터베이스별 검색 건수는 Fig. 1의 확인(Identification) 단계에 명시하였다.

4. 선정 및 제외 기준

1) 선정 기준

(1) 의학·치의학·간호학·약학·수의학 등 보건의료계열 선발에서 MMI를 사용한 연구, (2) 신뢰도·타당도·예측타당도·공정성·수용성·실행 가능성 중 하나 이상을 실증적으로 검토한 연구, (3) 스테이션이 윤리·인성·도덕성·공감·의사소통·전문직업성 등 윤리적 판단 및 인성·도덕성 관련 수행구인을 포함한 연구, (4) 비대면 MMI 또는 생성형 AI 문항 개발·평가 연구를 포함하였다. 이 선정 기준은 임의적 목록이 아니라 Table 1의 조작적 정의와 세 연구 질문에 대응하도록 구성한 것이다. 즉 기준 (1)·(2)는 RQ2의 심리측정 근거를, 기준 (3)은 RQ1의 윤리적 판단 및 인성·도덕성 관련 수행의 범주화를, 기준 (4)는 RQ3의 기술 환경 변화를 각각 포착하기 위한 이론적 근거를 갖는다. 문헌 유형별 포함의 근거는 다음과 같다. 1차 실증연구는 본 고찰의 핵심 분석 대상이고, 선행 체계적 문헌고찰 11편은 본 고찰의 종합 결과를 기존 종합과 대조·검증하기 위한 대조 근거로, 문항 구성 지침 1편[14]은 1차 실증연구는 아니나 RQ3의 설계 논의에 필요한 준거 문헌으로 포함하였다.

2) 제외 기준

비동료심사 입시 안내 자료·경험담, MMI가 아닌 일반 구조화 면접만 다룬 연구, 보건의료 선발과 무관한 기업 채용 연구, 방법·결과 정보가 부족한 초록·의견문은 제외하되, 국

www.kci.go.kr

Table 1. Operational definitions of key constructs in this review

Construct	Operational definition in this review
Character formation	Character is understood as developing over time rather than as a completed attribute at admission. It is not treated as a directly measurable construct in this review; only situation-specific judgments and behaviors relevant to character and morality are considered observable.
Character traits	Situation-specific behaviors and language associated with traits such as honesty, responsibility, and consideration. These observations are not interpreted as direct measures of stable character traits.
Morality	Situation-specific ethical judgment and moral sensitivity observable in responses and behavior. These observations are not interpreted as direct measures of moral character as a whole.
Professionalism	Values and behavioral norms inherent in the social contract of the profession, such as honesty, responsibility, altruism, and respect for patient autonomy.
Empathy	The capacity to grasp others' emotions and perspectives cognitively and affectively and to respond supportively. It includes non-verbal expression.
Virtue	A stable excellence of character exercised toward the ends of medicine beyond rule-following (e.g., benevolence, practical wisdom, and courage; [6]).

내 구조화 면접 타당도 논의에 기여하는 연구는 배경문헌으로 별도 표시하였다.

5. 자료 추출과 합성 방법

자료 추출 항목은 저자·연도·국가·계열·설계·표본·스테이션 수와 형식·평가 구인·신뢰도 지표·타당도 근거·예측 준거·온라인/AI 적용·주요 결과·한계로 구성하였다. RQ1 (질적)은 스테이션 내용을 귀납적으로 부호화한 뒤, 도출된 범주와 의료윤리 원칙 및 관련 전문직 가치 사이의 관계를 해석적으로 검토하는 내용분석으로 수행하고, 결과별 확신도를 CERQual의 네 요소(방법론적 한계·일관성·자료 적정성·관련성)로 평정하였다. RQ2·RQ3 (양적)은 결과 지표가 이질적이어서 SWiM 방향성 투표를 적용하여 각 결과의 효과 방향을 긍정(+)·변화없음(0)·혼재(±)로 분류하고, 해당 결과를 보고한 연구가 3편 미만이어서 방향 판정이 불가한 경우 '판정 보류'로 표기하였다. 3편 이상에서 70% 이상 방향이 일치하면 '일관'으로 종합하였으며 연구는 방법론적 질과 표본 규모 순으로 정렬하였다. 다만 방향성 투표의 '일관' 판정은 효과의 방향에 대한 종합일 뿐 효과의 크기나 절대 수준에 대한 종합이 아니므로, 결과 보고에서는 방향의 일관성과 수준의 변산을 구분하여 서술한다. 윤리적 원칙 및 전문직 가치와의 대응관계 해석에는 연구자의 해석이 개입되므로 두 부호화자(인격교육·의학교육 배경)가 독립 부호화 후 합의하고 상충 시 제3자가 중재하였으며, 부호화 지침과 합의 로그는 교신저자에게 요청 시 제공한다.

6. 방법론적 질 평가

포함 1차 연구는 MERSQI (6도메인 10문항, 5~18점)와 비무작위·관찰 설계에 대한 NOS-E (최대 6점)로 평가하였다[21,22]. 두 도구를 병용한 이유는 다음과 같다. MERSQI는 연구 설계·표본·자료 유형·타당도 근거·자료분석·결과 수준의 여섯 도메인으로 의학교육 연구 전반의 방법론적 질을 평가하는 데 강점이 있으나, 본 고찰의 1차 연구 대부분을 차지하는 비무작위·관찰 설계 특유의 비뚤림 위험(표본의 대표성, 비교집단의 구성, 결과 평가의 객관성 등)을 세분하여 포착하는 데에는 한계가 있다. NOS-E는 바로 이 관찰 설계의 비뚤림 위험 도메인을 보완적으로 평가하는 도구로서, 두 도구의 병용은 단일 도구 사용 시 발생하는 평가의 사각지대를 상호 보완하기 위한 것으로, MERSQI와 NOS-E를 상호 보완적 도구로 제시한 Cook과 Reed [22]를 따랐다. 대조 근거로 포함한 선행 체계적 문헌고찰 11편은 AMSTAR-2 [23]로 평가하였다. 두 평가자가 독립 채점하고 불일치는 협의로 해소하였으며, 총점의 단순 평균보다 타당도 근거와 결과 수준의 취약성을 중심으로 해석하였다. 동일 코호트에서 파생된 McMaster 계열과 국내 강원대 계열 연구는 근거의 이중 계산을 피하기 위해 코호트 단위로 묶어 서술하였다.

연구 결과

1. 문헌 선별 결과

6개 데이터베이스에서 총 952건을 확인하였다(PubMed 345, Embase 306, ERIC 19, RISS 270, KCI 7, DBpia 5). 중복 제거와 제목·초록 선별은 단일 통합 단계로 수행되었으며, 이 단계에서 중복 및 비관련 레코드 906건을 제외하고 46편을 전문 검토 대상으로 확정하였다. 전문 검토에서 MMI

www.kci.go.kr

Table 2. Evidence map of the included publications by research question

Theme (RQ)	Principal included publications	n
Foundational and predictive validity (RQ2)	Eva et al. (2004a) [7], (2004b) [24], (2009) [25], (2012) [26], Reiter et al. (2007) [27]	5
Korean empirical studies (RQ1, RQ2)	Roh et al. (2009) [8], Kim et al. (2014) [28], Kim et al. (2014) [29], Lee et al. (2016) [30], Kim et al. (2017) [31], Kim et al. (2021) [32]	6
Constructs, personality, and fairness (RQ1, RQ2)	Griffin and Wilson (2012) [33], Jerant et al. (2012) [34], McLaughlin et al. (2017) [35], Breil et al. (2020) [36], Renaud et al. (2022) [37], Ayub et al. (2024) [38], Szkwara et al. (2025) [39], Góes et al. (2025) [40]	8
Extension to other health professions (RQ1)	Duff et al. (2020) [41]	1
Remote and online formats (RQ3)	Kok et al. (2021) [42], Carroll Turpin et al. (2021) [43], Domes et al. (2021) [44], Inzana et al. (2022) [45], Sabesan et al. (2022) [46], Hammond et al. (2023) [47], Callwood et al. (2023) [48], Cork (2023) [49], Iramaneerat et al. (2025) [50], Auton et al. (2025) [51], Sa et al. (2025) [52], Yolanda et al. (2020) [53]	12
Generative artificial intelligence (RQ3)	Hamila et al. (2026) [13], Menon et al. (2026) [14]	2
Prior systematic reviews (comparison)	Pau et al. (2013) [9], Rees et al. (2016) [10], Patterson et al. (2016) [12], Callwood et al. (2018) [11], Kelly et al. (2018) [54], Roberts et al. (2018) [55], Yusoff (2019) [56], Lin et al. (2022) [57], Wildermuth (2022) [58], Huang et al. (2024) [59], Lin et al. (2024) [60]	11
Total		45

Note. In the generative artificial intelligence row, Hamila et al. (2026) is a primary empirical study, whereas Menon et al. (2026) is a station-construction guide. RQ, research question.

특이적이지 않은 구조화면접 연구 1편을 제외하여 최종 45편(1차 실증연구 33편, 선행 체계적 문헌고찰 11편, 문항 구성 지침 1편)을 포함하였다(Fig. 1).

2. 포함 연구의 특성

포함 연구는 캐나다·한국·미국·호주·영국·독일·태국·브라질·인도네시아·파키스탄·브루나이 등 다국적이며 의학이 다수이되 치의학·수의학·약학·간호학을 포괄하였다. 연구 질문별 근거 지도는 Table 2에, 1차 연구의 상세 특성은 Table 3에 제시하였다.

3. 방법론적 질 평가 결과

1차 연구 33편의 MERSQI 총점 평균은 약 12.4점(18점 만점; 범위 약 10~15)으로 MERSQI를 적용한 선행 연구 26편의 중앙값(11.3점)을 상회하였고[22], NOS-E 중앙값은 약 3.3점(6점 만점)이었다. 두 평가자의 총점 급내상관계수(ICC)는 0.82, 도메인 수준 가중 카파(κ)는 0.71~0.90으로 양호하였다. 도메인별로는 '타당도 근거'와 '결과(outcomes)'에서 낮은 점수가 반복되어, 다수 연구가 낮은 수준의 교육성과를 평가하거나 단일기관 설계를 사용함으로써 외적 타당성이 제한되는 경향을 보였다. 선행 체계적 문헌고찰 11편의 AMSTAR-2 평가에서는 프로토콜 사전 등록과 출판편향 평가 항목에서 감점되는 경향이 있었다.

4. RQ1: MMI 스테이션에서 관찰되는 윤리적 판단과 인성·도덕성 관련 수행

MMI 스테이션은 추상적인 인성이나 도덕성을 직접 묻기보다, 구체적인 상황에서 나타나는 판단·의사소통·행동을 관찰하도록 구성되어 있었다. 문헌에 제시된 스테이션의 내용과 평가 영역을 종합한 결과, 윤리적 딜레마, 대인관계에서의 진실성과 책임, 공감적 의사소통, 공적 책임의 네 범주가 두드러졌으며(Table 4), 이 범주화는 국가와 보건의료 직종 전반에서 일관되게 확인되어 확신도 '높음'으로 평가되었다(Table 5). 보건의료계열 MMI의 개인적 평가 영역을 종합한 선행 체계적 문헌고찰에서도 의사소통, 팀워크·협력, 윤리적·도덕적 판단을 포함한 다양한 영역이 반복적으로 확인되었다[11]. 한편 구인타당도 연구에서는 MMI 점수가 의도한 개별 구인보다 스테이션 특이적 수행을 주로 반영하거나[36], 평가자의 엄격성이 점수 변동의 상당 부분을 설명하는 것으로 보고되었다[35]. 스테이션별로 구분되는 다차원 요인구조를 지지한 연구도 있었으나, 이 경우에도 각 구인이 하나의 스테이션에만 대응되어 구인과 스테이션의 효과를 분리

www.kci.go.kr

Table 3. Characteristics of 21 representative included publications

Author (year)	Country, field, design	Sample and stations	Constructs, reliability/validity	Main findings
Eva et al. (2004)	Canada, medicine, validation	Multiple stations, multiple raters	General non-cognitive; generalizability theory	Prototype of the MMI; more stable than individual interviews
Reiter et al. (2007)	Canada, medicine, longitudinal	117 (45 admitted)	Clinical and ethical domains (CLEO/PHELO); predictive validity	Best predictor of OSCE and clerkship performance; significantly predicted MCCQE Part I CLEO/PHELO scores
Eva et al. (2012)	Canada, medicine, large-scale longitudinal	National cohort	Professional domains; criterion validity	Associated with national licensing examination scores
Roh et al. (2009)	Korea, medicine, cross-sectional	84; 3 stations × 8 min	Motivation, interpersonal communication, ethical decision-making; G = 0.791	First in Korea; high acceptability
Kim et al. (2014)	Korea, medicine, cohort	Graduate-entry students (2008 entrants)	Non-cognitive; predictive validity	Moderate correlations with medical humanities and clinical practice courses
Kim et al. (2014)	Korea, medicine, comparative	First-year students by admission track	Non-cognitive; group comparison	Differences in characteristics of students admitted through the MMI track
Lee et al. (2016)	Korea, medicine, longitudinal	Kangwon National University	Critical thinking, professionalism, communication; predictive validity	Predicted achievement in the first 2 years
Kim et al. (2017)	Korea, medicine, case study	Single institution	Operation; feasibility	Reported utility and operational experience
Kim et al. (2021)	Korea, medicine, validity	Video-based stations	Empathy; feasibility	Video-based assessment of empathy was useful.
Griffin and Wilson (2012)	Australia, medicine, correlational	Applicants	Big Five personality; construct	MMI scores were associated with extraversion and conscientiousness across all 3 years and with agreeableness in 2 years.
Jerant et al. (2012)	USA, medicine, cohort	Applicants	Personality, admission offer; construct	Extraversion was associated with MMI performance; extraversion and agreeableness were associated with acceptance offers.
McLaughlin et al. (2017)	USA, pharmacy, Rasch analysis	Stations targeting specific constructs	Constructs; internal structure	Rater stringency explained a substantial portion of score variance
Breil et al. (2020)	Germany, medicine, Bayesian generalizability theory	Stations, skills, raters	Constructs; construct validity	Reflected the overall quality of situational performance

www.kci.go.kr

Table 3. Continued

Author (year)	Country, field, design	Sample and stations	Constructs, reliability/validity	Main findings
Renaud et al. (2022)	Canada, medicine, experimental	Station format manipulated	Psychometric properties	Format affected reliability
Ayub et al. (2024)	Pakistan, medicine, factor analysis	Multiple stations	Internal structure; construct	Internal structure verified
Szkwara et al. (2025)	Australia, medicine, measurement equivalence	Group comparison	Fairness; measurement invariance	Assessed equity between groups
Iramaneerat et al. (2025)	Thailand, medicine, retrospective comparison	340; 4 onsite / 4 online administrations (8~12 stations each)	Mean α 0.49 / 0.51; validity	Reliability comparable between formats, but low in both
Kok et al. (2021)	Brunei, medicine, implementation	Video-conference MMI	Non-cognitive; feasibility	MMI could be conducted during the COVID-19 pandemic
Hammond et al. (2023)	USA, pharmacy, validity	7 virtual stations	Integrity, empathy, critical thinking, teamwork, adaptability; validity	Validity evidence for a virtual MMI
Hamila et al. (2026)	Australia, medicine, psychometric	17 AI-generated vs 39 conventional stations	7 character-trait domains; internal consistency	A higher proportion of AI-generated stations reached $\alpha > 0.85$
Menon et al. (2026)	Multinational, health professions, station-construction guide	Design principles	AI-resilient design	Proposed principles for interview design in the era of AI

Note. Twenty-one representative publications are presented owing to space limitations. Menon et al. (2026) is a station-construction guide rather than a primary empirical study but is presented together because it serves as a reference for the discussion of research question 3.
α, Cronbach's alpha; AI, artificial intelligence; CLEO/PHELO, Population Health and the Considerations of the Legal, Ethical and Organisational Aspects of Medicine (subsections of the MCCQE Part I); COVID-19, coronavirus disease 2019; G, generalizability coefficient; MCCQE, Medical Council of Canada Qualifying Examination; MMI, multiple mini-interview; OSCE, objective structured clinical examination.

하기는 어려웠다[38]. 따라서 MMI 점수를 특정 덕목이나 성격 특성의 독립적인 측정치로 해석하는 데에는 주의가 필요하다.

5. RQ2: 신뢰도·타당도 근거와 수행 해석의 한계

MMI의 신뢰도·타당도 근거를 영역별로 종합한 결과와 그 해석의 한계는 Table 6에 제시하였다. MMI의 신뢰도는 스테이션 수, 평가자 훈련, 채점 루브릭 등 평가 설계에 따라 크게 달랐다. 포함 연구에서 Cronbach α는 약 0.34~0.98, 일반화가능도 계수(G)는 최대 0.88까지 보고되었다. 따라서 MMI라는 형식 자체가 높은 신뢰도를 보장한다고 보기는 어렵다. 다만 여러 스테이션과 평가자를 통해 수행을 반복적으로 표집하는 구조는 단일 면접보다 안정적인 평가를 가능하게 하는 방향으로 보고되어 왔다[7,9,10]. Iramaneerat 등[50]은 대면 4회와 온라인 4회의 MMI (회당 8~12개 스테이션)를 비교하여 평균 신뢰도가 각각 $\alpha = 0.49$와 0.51로 유사했지만, 두 방식 모두 절대적인 신뢰도 수준은 낮았다. 국내 연구에서는 3개 스테이션으로 운영한 MMI에서 일반화가능도 계수 0.791과 높은 수용성이 보고되었다[8]. 이러한 결과는 MMI의 신뢰도를 평가할 때 면접 형식뿐 아니라 충분한 수행 표집과 평가 설계를 함께 고려해야 함을 보여준다.

타당도 측면에서도 MMI 점수를 특정한 성격 특성이나 덕목의 직접적인 측정치로 해석하기는 어렵다. 성격 5요인과 MMI 수행의 관계를 분석한 연구들은 서로 다른 양상을 보였다. Griffin과 Wilson [33]은 외향성과 성실성, 일부 연도에서는 친화성과 MMI 점수의 관련성을 보고한 반면, Jerant 등[34]에서는 외향성이 MMI 수행과 유의한 관련을 보였고, Góes 등[40]에서는 성실성만이 유의하게 관련되었다. 이러한 결과는 MMI 수행과 성격 특성 사이에 일부 관련성이 존재하더라도, MMI 점수를 하나의 고정된 성격 특성으로 환원하여 해석하기 어렵다는 점을 보여준다. 예측타당도에서는 Reiter 등[27]이 MMI 점수와 이후 임상실습 평가, OSCE 및 국가면허시험의 일부 영역 사이의 유의한 관련성을 보고하

www.kci.go.kr

Table 4. Interpretive mapping of multiple mini-interview station types to ethical principles and professional values

Station type	Related ethical principles and professional values	Representative scenarios	Indicators of excellent performance
Ethical dilemmas and value conflicts	Respect for autonomy, beneficence, non-maleficence, justice, practical wisdom	Refusal of blood transfusion; confidentiality versus prevention of harm; conflict between rules and the protection of life	Identifies conflicting values, explains the perspectives of those involved empathically, and presents a judgment that considers both principles and context
Interpersonal ethics and integrity	Honesty, responsibility, moral courage, sense of community	A friend's misconduct; free-riding in group work; witnessing a colleague's unprofessional behavior	Goes beyond the dichotomy of reporting versus tacit acceptance by proposing a balanced, stepwise response that includes communication, an opportunity for reflection, and protection of community norms
Empathy and patient-centered communication	Compassion, respect, rapport, sensitivity to vulnerability	Delivering bad news; a patient's anger or anxiety; the needs of a vulnerable patient	Checks emotions and uses supportive language before conveying information, and respects the patient's level of understanding and right to choose
Public responsibility and health policy	Distributive justice, social responsibility, public interest, equity	Pricing of orphan drugs; vaccine-refusing communities; allocation of medical resources	Balances individual liberty and public safety, incentives for innovation and access, and protection of the majority and vulnerable groups

Table 5. Confidence in the qualitative findings for research question 1

Review finding (summary)	Confidence	Main evidence and judgment
Stations are categorized into four types: ethical dilemmas, integrity, empathy, and public responsibility	High	Consistent across countries and professions; adequate data
Each type can be interpreted in relation to ethical principles and professional values such as autonomy, honesty, compassion, and public interest.	Moderate	Partly downgraded for relevance owing to interpretive involvement
Individual constructs were not consistently separable as independent dimensions.	Moderate	Supported by Breil et al. (station-specific variance) and McLaughlin et al. (rater effects); Ayub et al. supported a station-level multidimensional structure in which stations and constructs were confounded; downgraded for coherence

Note. Confidence was assessed using the GRADE-CERQual approach [20]. CERQual, Confidence in the Evidence from Reviews of Qualitative Research; GRADE, Grading of Recommendations Assessment, Development and Evaluation.

였다. Eva 등[26]의 대규모 코호트에서도 MMI를 포함한 선발 절차를 통해 합격한 지원자들이 다른 기관에 진학한 불합격 지원자보다 캐나다 국가면허시험에서 높은 성적을 보였다. 국내 연구에서도 MMI 점수와 일부 의학교육 과정의 학업성취 및 수행평가 사이의 관련성이 보고되었다[29,30]. 그러나 지금까지 확인된 예측 근거의 중심은 학업성취, 임상실습, OSCE, 국가면허시험과 같은 교육·평가 성과에 있다. 입학 단계에서 관찰된 윤리적 판단이나 관련 수행이 실제 임상 현장에서 지속적인 윤리적 행동으로 이어지는지를 보여주는 장기 종단 근거는 여전히 충분하지 않다[9,10]. SWiM 방향성 투표 결과, 신뢰도·예측타당도·구인타당도는 일관된 방향으로, 공정성은 혼재된 방향으로 종합되었다. 수용성은 대체

www.kci.go.kr

Table 6. Synthesis of evidence on the reliability and validity of the multiple mini-interview and limits of interpretation

Domain of evidence	Synthesized finding	Direction (SWiM)	Interpretation and limitations
Reliability	Reliability varied widely across studies (α approximately 0.34~0.98) and was influenced by assessment design, including the number of stations, scoring rubrics, and rater training. Multi-station, multi-rater designs were generally more stable than single-interview approaches.	+ (consistent; direction only)	The MMI format alone does not guarantee high reliability; adequate sampling across stations and raters and careful assessment design are required.
Predictive validity	Associations were reported with subsequent educational and assessment outcomes, including academic achievement, clerkship performance, OSCE performance, and licensing examinations; the strength of these associations varied across studies.	+ (consistent)	Evidence supports associations with selected subsequent educational and assessment outcomes, but does not establish sustained ethical behavior in clinical practice.
Construct validity	Studies of internal structure and personality associations did not show a simple one-to-one correspondence between MMI scores and discrete traits; performance was influenced by multiple sources, including stations, assessed domains, and raters.	+ (consistent)	MMI scores should not be interpreted as direct or independent measures of specific virtues or stable personality traits.
Fairness and acceptability	Studies generally reported acceptable feasibility and participant acceptance, while findings related to fairness varied across demographic, cultural, linguistic, and technological contexts.	± (mixed)	Fairness should be evaluated empirically across applicant groups and delivery contexts rather than assumed from the structured format alone.
Long-term clinical ethical behavior	Evidence directly linking admission-stage MMI performance with sustained ethical or professional behavior in clinical practice was insufficient.	Not rated (insufficient evidence)	Current evidence does not support inferring enduring moral character or long-term ethical behavior from MMI performance at admission.

Note. The direction of effect was classified by vote counting according to the SWiM guideline [19] as positive (+), no change (0), or mixed (±). A finding was synthesized as consistent when at least 70% of three or more studies agreed in direction and was recorded as not rated when fewer than three studies reported it. α, Cronbach's alpha; MMI, multiple mini-interview; OSCE, objective structured clinical examination; SWiM, Synthesis Without Meta-analysis.

로 긍정적이었으며, 장기적 임상윤리 행동은 근거 부족으로 판정하지 않았다(Table 6).

6. RQ3: 비대면 MMI와 생성형 인공지능의 윤리적·방법론적 쟁점

COVID-19 이후 비대면 MMI가 여러 보건의료계열 선발에 도입되었으며, 포함된 12편의 관련 연구에서는 온라인 환경에서도 MMI의 실행 가능성과 수용성이 대체로 확인되었다[42-44]. 다만 신뢰도는 연구와 설계에 따라 차이가 있었다. Iramaneerat 등[50]은 대면 4회와 온라인 4회의 MMI (회당 8~12개 스테이션)에서 평균 신뢰도가 각각 α=0.49와 0.51로 유사하다고 보고했지만, 두 값 모두 높은 수준은 아니었다. 약학과 수의학을 포함한 다른 보건의료계열에서도 가상 MMI의 타당도 근거와 실행 가능성이 보고되었으며[45,47], 비동시적 MMI 역시 실제 선발에 적용될 수 있음이 확인되었다[48]. 이러한 결과는 비대면 MMI가 실행 가능한 대안이 될 수 있음을 보여주지만, 평가의 질이 온라인 형식 자체만으로 보장되는 것은 아니며 스테이션 구성, 평가자, 기술 환경과 접근성 등을 함께 고려해야 함을 보여준다.

생성형 인공지능과 관련해서는 1편의 실증연구와 1편의 문항 구성 지침이 확인되었다. Hamila 등[13]은 생성형 인공지능으로 개발한 17개 스테이션과 기존 방식으로 개발한 39개 스테이션을 비교하였으며, 두 유형의 문항이 전반적으로 유사한 심리측정적 성능을 보였고 AI 생성 문항 가운데 높은 내적일관성을 보인 문항의 비율도 높았다고 보고하였다. 그러나 이는 생성형 인공지능이 MMI 문항 개발에서 활용 가능성을 보였다는 초기 근거이며, 공정성이나 장기적인 타당성까지 입증한 결과는 아니다. Menon 등[14]은 생성형 인공지능이 지원자의 응답 생성에 개입할 수 있는 환경에서 탐지 중심의 대응보다 평가하려는 구인을 명확히 하고 면접 자체를 이에 맞게 설계하는 접근이 필요하다고 제안하였다. 따라서 현재의 근거는 생성형 인공지능의 활용 가능성을 보여주는 동시에, 구인 명료화, 공정성, 응답의 진정성, 인간의 검토와 책임에 관한 추가 검증이 필요함을 보여준다.

www.kci.go.kr

논 의

1. 상황 속 수행의 관찰과 해석의 범위

이 고찰에서 MMI 스테이션은 윤리적 딜레마, 대인관계에서의 진실성과 책임, 공감적 의사소통, 공적 책임과 같이 구체적인 상황 속에서 나타나는 판단과 행동을 관찰하도록 구성되어 있었다. 이러한 결과는 MMI가 지원자의 인격이나 도덕성을 직접 측정하기보다, 특정한 상황에서 드러나는 윤리적 판단과 인성·도덕성 관련 수행을 표집하는 도구라는 해석을 뒷받침한다. 실제로 구인타당도 연구에서는 개별 평가 영역이 항상 뚜렷하게 분리되지 않았고, 수행은 스테이션과 평가영역, 평가자의 영향을 함께 받는 것으로 나타났다[35,36]. 스테이션 수준의 다차원 구조가 지지된 경우에도 구인과 스테이션이 일대일로 대응되어 있어 두 효과를 구분하기 어려웠다[38]. 따라서 하나의 MMI 점수나 특정 스테이션의 수행을 정직성·공감·책임과 같은 안정된 개인 특성의 직접적인 측정치로 해석하는 데에는 주의가 필요하다.

그렇다고 이러한 구인 미분화를 곧바로 MMI가 '통합된 인격'이나 '실천적 지혜'를 측정한다는 증거로 해석할 수도 없다. 서로 다른 구인이 분리되지 않는 데에는 실제 수행의 복합성뿐 아니라 스테이션 특이성, 평가자 효과, 후광 효과와 같은 측정상의 요인도 작용할 수 있기 때문이다. 오히려 이 결과가 분명하게 보여주는 것은 윤리적 판단과 대인관계 수행이 맥락과 분리된 고정 속성으로 쉽게 환원되지 않는다는 점이다. 이러한 관점에서 MMI의 가치는 사람됨 전체를 판정하는 데 있기보다, 여러 상황과 평가자를 통해 관련 수행의 표본을 넓히고 그 범위 안에서 제한적으로 판단할 수 있게 하는 데 있다.

2. 의료윤리 원칙과 상황적 판단의 관계

RQ1에서 확인된 네 범주는 자율성 존중, 선행, 악행금지, 정의와 같은 생명의료윤리의 원칙과 관련하여 해석할 수 있다[5]. 윤리적 딜레마와 공적 책임을 다루는 스테이션에서는 서로 충돌하는 가치와 이해관계를 어떻게 파악하고 조정하는지가 드러나며, 대인관계와 공감적 의사소통을 다루는 스테이션에서는 타인의 관점과 취약성을 고려하면서 자신의 판단을 어떻게 표현하는지가 관찰된다. 그러나 이러한 대응관계가 각 스테이션이 특정한 윤리 원칙이나 덕목을 직접 측정한다는 의미는 아니다. 하나의 상황에서도 여러 가치가 동시에 작용하며, MMI가 제공하는 것은 그 상황에서 나타난 판단과 의사소통의 수행 표본이다.

덕 윤리는 이러한 상황적 판단을 이해하는 데 보완적인 관점을 제공한다. 좋은 의료실천은 원칙을 적용하는 데 그치지 않고, 환자와 상황의 구체성을 고려하여 책임 있게 판단하고 행위하는 실천적 지혜와 관련된다[6]. 그러나 MMI에서 관찰된 수행을 곧 실천적 지혜나 덕의 소유로 동일시할 수는 없다. 선발 장면에서 확인되는 것은 특정 조건에서 나타난 판단과 행동이며, 그러한 수행이 다른 상황과 실제 임상에서도 지속되는지는 별도의 경험적 검증이 필요하다[9,10]. 따라서 생명의료윤리의 원칙과 덕 윤리는 MMI가 사람의 도덕성을 판정하도록 정당화하는 근거라기보다, 관찰된 수행이 지닌 의미와 한계를 해석하는 이론적 틀로 사용하는 것이 적절하다.

3. 선발과 교육의 연속성: 판단의 범위와 이후의 형성

예측타당도에 관한 근거가 보여주는 것은 선발과 이후 교육 사이의 일정한 연관성과, 동시에 그 해석의 경계이다. MMI 수행은 일부 학업성취, 임상실습, OSCE, 국가면허시험과 관련되어 있었지만, 입학 단계에서 관찰된 윤리적 판단과 관련 수행이 수년 뒤 임상현장의 지속적인 윤리적 행동으로 이어지는지를 보여주는 근거는 충분하지 않았다[9,10]. 따라서 선발에서 얻은 수행 표본은 한 사람의 이후 모습을 미리 확정하는 자료가 아니라, 특정한 시점과 조건에서 관찰된 제한된 정보로 이해해야 한다.

이 경계를 분명히 할 때 선발과 교육의 역할도 선명해진다. 선발은 지원자가 주어진 상황에서 무엇을 인식하고, 어떻게 판단하며, 타인에게 어떻게 응답하는지를 관찰할 수 있다. 그러나 의료인에게 요구되는 윤리적 판단, 책임, 공감, 전문직업성은 입학 시점에 완결되어 있는 자질이라고 보기 어렵다. 이러한 역량과 태도는 교육과 임상 경험, 관계와 성찰을 거치며 계속 형성되고 변화한다. 그러므로 선발의 타당성은 '완성된 사람'을 가려냈는가에 있지 않고, 제한된 조건에서 나타난 수행을 얼마나 공정하고 신뢰롭게 관찰했으며 그 결과를 어디까지 절제하여 해석했는가에 달려 있다.

따라서 선발과 입학 후 교육은 역할이 다르지만 서로 분리된 과정으로 볼 필요는 없다. 선발에서 중요하게 다루는 윤리적 판단, 타인에 대한 존중, 공감적 의사소통, 공적 책임이 입학 후 의료교육에서도 지속적으로 다루어야 할 가치라면, 두 단계 사이에는 교육적 정합성이 필요하다. 다만 선발 점수를 곧바로 입학 후 교육의 진단자료로 사용하는 방식은 아직 충분히 검증되지 않았다. 앞으로는 입학 당시 관찰된 수행과 이후의 의료윤리·전문직업성·임상수행을 장기간 연결하여 추적함으로써, 선발에서 포착된 수행이 무엇을 예측하고 무엇은 예측하지 못하는지를 경험적으로 밝혀야 한다. 이러한 선발과 입학 후 교육의 관계를 Fig. 2에 개념적으로 제시하였다.

4. 기술 환경의 변화와 평가의 책임

비대면 MMI와 생성형 인공지능의 도입은 평가의 방식을 바꾸었지만, 무엇을 평가하고 그 결과를 어디까지 신뢰

www.kci.go.kr

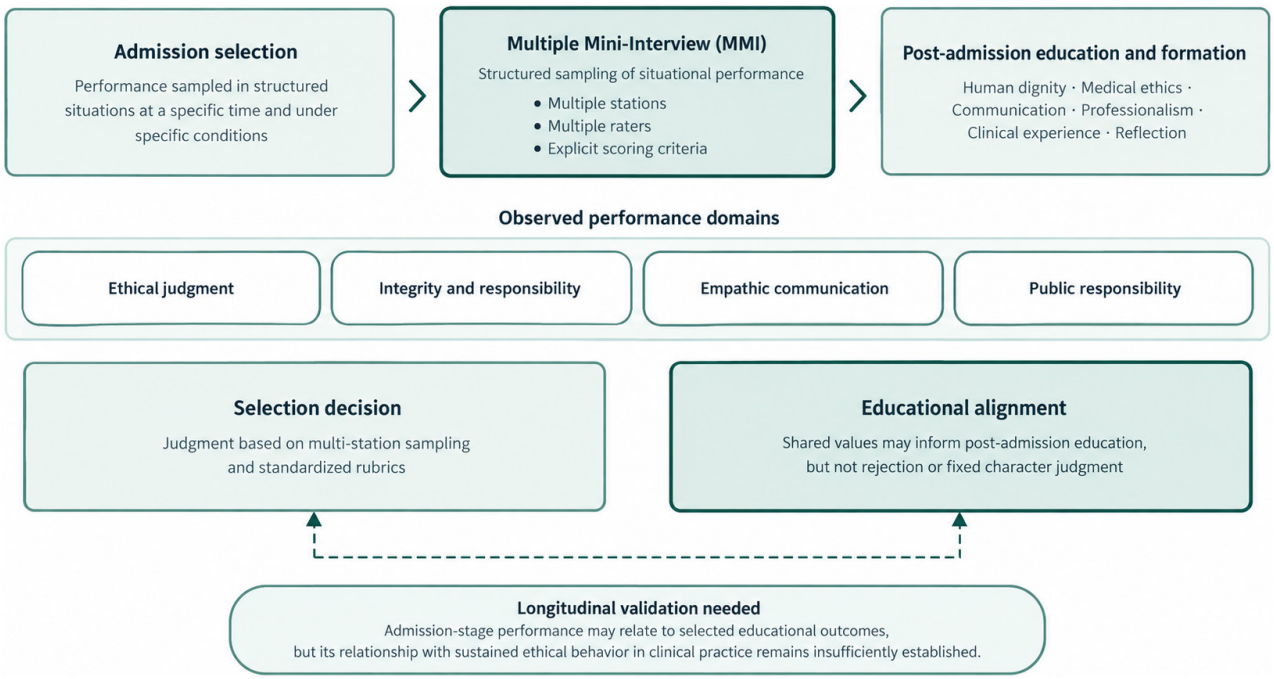


Fig. 2. Conceptual framework illustrating educational alignment between performance sampled at admission and continued ethical and professional formation after admission. MMI, multiple mini-interview.

할 것인가라는 문제를 없애지는 않았다. 비대면 MMI는 여러 보건의료계열에서 실행 가능성과 수용성을 보여주었으며[42-45,47,48,50], 대면과 온라인 환경에서 유사한 신뢰도가 보고된 연구도 있었다[50]. 그러나 평가의 질이나 공정성이 온라인이라는 형식만으로 보장되는 것은 아니다. 집단 간 측정 동등성을 검토한 연구 역시 동일한 평가가 서로 다른 지원자 집단에서 같은 의미를 갖는지 경험적으로 확인해야 함을 보여준다[39]. 따라서 기술적 접근성, 언어와 문화, 평가 환경의 차이가 수행과 점수에 미치는 영향은 지속적으로 검토할 필요가 있다.

생성형 인공지능에서도 같은 원칙이 적용된다. AI로 개발한 MMI 스테이션이 기존 방식으로 개발한 스테이션과 유사한 심리측정적 성능을 보였다는 초기 근거는 문항 개발의 새로운 가능성을 보여준다[13]. 그러나 문항이 안정적으로 작동한다는 것과 그 문항이 평가하려는 윤리적·전문직적 수행을 적절하게 표집한다는 것은 같은 문제가 아니다. 또한 지원자가 생성형 인공지능을 활용할 수 있는 환경에서는 AI 사용을 탐지하는 데만 초점을 맞추기보다, 평가하려는 구인을 먼저 명료하게 하고 그 구인이 실제 수행에서 드러나도록 면접을 설계할 필요가 있다[14]. 결국 기술은 평가의 효율성을 높일 수 있지만, 무엇을 관찰할 것인지, 어떤 수행을 의미 있게 해석할 것인지, 그 결과로 어디까지 판단할 것인지에 대한 책임까지 대신할 수는 없다.

따라서 기술 기반 MMI에서는 평가 구인의 명료화, 집단 간 공정성의 검증, 인간의 검토와 책임, 그리고 결과 해석의 절제가 함께 요구된다.

5. 선발과 연구를 위한 함의

실제 선발에서는 MMI라는 형식을 도입하는 것만으로 충분하지 않다. 이 고찰에서 신뢰도는 스테이션의 수와 구성, 평가자, 채점 기준 등 평가 설계에 따라 크게 달랐다. 따라서 먼저 무엇을 관찰하려는지를 분명히 하고, 해당 수행을 여러 상황과 평가자에 걸쳐 충분히 표집할 수 있도록 면접을 설계해야 한다. 특히 윤리적 판단, 공감, 책임과 같이 맥락의 영향을 받는 수행을 하나의 스테이션이나 단일 점수로 환원하여 해석해서는 안 된다. 비대면 또는 AI 기반 방식을 사용할 때에도 기술적 편의보다 구인타당도, 공정성, 접근성과 인간의 책임을 우선하여 검토할 필요가 있겠다.

향후 연구에서 가장 중요한 과제는 입학 단계에서 관찰된 수행과 이후의 실제 임상행동을 장기간 연결하는 것이다. 지금까지의 예측 근거는 학업성취, 임상실습, OSCE, 국가면허시험과 같은 비교적 가까운 준거에 집중되어 있다[9,10]. 앞으로는 다기관 종단 연구를 통해 입학 당시의 윤리적 판단과 의사소통 수행이 이후의 환자 중심 행동, 전문직업성, 윤리적 의사결정과 어떠한 관계를 갖는지 검증할 필요가 있다. 동시에 그러한 관계가 입학 당시의 개인차에서 비롯된 것인지, 교

www.kci.go.kr

육과 임상경험을 거치며 형성되고 변화한 결과인지를 구분할 수 있는 연구 설계가 필요하다고 여겨진다.

6. 강점과 한계

이 고찰의 강점은 여섯 개 데이터베이스를 이용해 보건의료계열 MMI 문헌을 폭넓게 검색하고, 33편의 1차 실증연구를 핵심 근거로 분석하면서 11편의 선행 체계적 문헌고찰과 1편의 문항 구성 지침을 구분하여 해석했다는 데 있다. 1차 연구는 MERSQI와 NOS-E로, 선행 체계적 문헌고찰은 AMSTAR-2로 방법론적 질을 평가하였으며, 서로 이질적인 근거는 SWiM과 GRADE-CERQual을 이용하여 종합하였다. 또한 동일 코호트에서 파생된 연구를 구분하여 근거의 중복 계산을 줄이고, 윤리적 판단과 인성·도덕성 관련 수행을 인격 자체와 구별함으로써 MMI에서 실제로 관찰되는 것과 그 결과로부터 추론할 수 있는 범위를 분리하여 해석하고자 하였다.

한계도 분명하다. 첫째, 포함 연구의 MMI 설계와 결과지표가 이질적이어서 효과크기를 통합한 메타분석을 수행하지 못하였다. 둘째, 장기 예측타당도에 관한 근거가 부족하여 입학 단계의 수행과 실제 임상현장의 지속적인 윤리적 행동 사이의 관계를 판단하기 어렵다. 셋째, 스테이션의 내용을 의료윤리 원칙 및 관련 전문직 가치와 대응시키는 과정에는 연구자의 해석이 개입하므로, 이 고찰에서 제시한 범주와 대응관계가 유일한 분류라고 볼 수 없다. 넷째, 한국어와 영어 문헌으로 검색을 제한하고 회색문헌을 포함하지 않아 언어 및 출판편향의 가능성이 남아 있다. 마지막으로 비대면 MMI와 생성형 인공지능에 관한 근거는 아직 비교적 초기 단계이므로, 공정성·구인타당도·장기적 결과에 관한 해석에는 신중함이 필요하다.

결론 및 제언

이 연구는 보건의료계열 학생 선발에서 MMI가 무엇을 관찰할 수 있으며, 그 관찰 결과로부터 어디까지 말할 수 있는지를 검토하였다. 33편의 1차 실증연구를 중심으로 45편의 문헌을 종합한 결과, MMI 스테이션은 윤리적 딜레마, 대인관계에서의 진실성과 책임, 공감적 의사소통, 공적 책임이라는 네 범주의 수행을 주로 다루고 있었다. 복수의 스테이션과 평가자를 활용하는 구조는 전통적 면접보다 안정적인 평가를 가능하게 하는 방향의 근거를 보였으나, 실제 신뢰도는 평가 설계에 따라 크게 달라지는 것으로 나타났다. 또한 MMI 수행은 일부 학업성취, 임상실습, OSCE, 국가면허시험과 관련되어 있었지만, 입학 단계에서 관찰된 윤리적 판단과 관련 수행이 실제 임상현장의 지속적인 윤리적 행동으로 이어지는지를 보여주는 장기 근거는 충분하지 않았다.

이 결과가 허용하는 해석의 범위는 분명하다. MMI가 제공하는 것은 여러 구조화된 상황에서 드러난 판단·의사소통·행동의 표본이지, 지원자의 인격이나 도덕성 전체에 대한 판정이 아니다. 따라서 MMI의 타당성은 지원자의 인격을 얼마나 정확히 판정하는가가 아니라, 제한된 상황에서 나타난 수행을 얼마나 공정하고 신뢰롭게 관찰하고 그 결과를 얼마나 절제하여 해석하는가에서 찾아야 한다. 윤리적 판단, 책임, 공감과 전문직업성은 입학 시점에 완결된 자질이라기보다 이후의 교육과 임상 경험, 관계와 성찰 속에서 계속 형성되고 변화한다. 이는 선발의 역할을 축소한다는 뜻이 아니라, 선발에서 관찰된 수행의 의미를 그 범위 안에서 정확하게 이해하고, 그 이후의 형성을 교육의 과제로 이어가야 한다는 뜻이다.

그렇다면 입학 후 교육의 역할은 분명해진다. 선발에서 관찰된 수행을 고정된 개인 특성으로 받아들이는 것이 아니라, 학생이 다양한 윤리적 상황에서 무엇이 문제인지 알아차리고, 충돌하는 가치를 비교하며, 자신의 판단을 설명하고 다시 성찰하도록 돕는 것이다. 의료윤리와 전문직업성 교육도 원칙과 규범을 배우는 데서 그치지 않고, 실제 또는 모의 상황에서 판단하고 의사소통한 뒤 피드백과 성찰을 거치는 경험으로 이어질 필요가 있다. 이러한 변화의 과정을 선발 단계의 수행과 장기적으로 연결하여 살펴볼 때, 무엇이 입학 당시의 개인차이고 무엇이 교육과 임상경험을 통해 형성된 변화인지도 보다 분명하게 이해할 수 있을 것이다.

앞으로의 연구 과제는 선발과 교육 사이의 관계를 경험적으로 확인하는 데 있다. 입학 단계에서 관찰된 MMI 수행과 이후의 의료윤리·전문직업성·환자 중심 행동을 연결하는 다기관 종단 연구가 필요하며, 그 과정에서 교육과 임상 경험을 거치며 나타나는 변화도 함께 추적해야 한다. 그래야 선발에서 포착된 수행이 무엇을 예측하고 무엇은 예측하지 못하는지, 또 교육과 임상경험이 그 이후의 변화에 어떠한 역할을 하는지를 구분할 수 있다. 비대면 MMI와 생성형 인공지능의 활용에서도 같은 원칙이 적용된다. 기술적 가능성 자체보다 무엇을 평가하려는지, 서로 다른 지원자에게 공정하게 작동하는지, 그 결과를 어디까지 해석할 수 있는지, 그리고 그 판단에 대한 인간의 책임을 어떻게 유지할 것인지가 우선적으로 검증되어야 한다.

이 연구의 교육적 함의는 다시 해부학 실습실에서 선명해진다. 의학교육의 초기에 학생이 마주하는 것은 단순한 인체가 아니라 한 생을 살아낸 사람의 몸이다. 그 몸을 탐구의 대상으로 다루면서도 한 사람의 삶과 존엄을 잊지 않는 것이 해부학 교육의 윤리적 요청이라면, 선발에서도 지원자를 짧은 면접의 수행이나 하나의 점수로 환원하지 않는 같은 윤리

www.kci.go.kr

적 절제가 필요하다. MMI는 여러 상황에서 나타나는 판단과 태도를 구조화하여 관찰할 수 있지만, 그 수행이 한 사람의 인격에 대한 최종 결론이 될 수는 없다. 의료인의 윤리적 삶은 입학 순간에 완성되지 않으며, 교육과 경험 속에서 계속 형성된다. 결국 선발과 교육의 정합성은 동일한 평가를 반복하는 데 있지 않다. 선발에서는 한 사람을 짧은 수행이나 점수로 환원하지 않고, 교육에서는 그 사람이 더 나은 윤리적 판단과 실천을 형성해 갈 수 있도록 돕는 것, 바로 그 두 과정에서 인간존엄의 원칙을 일관되게 지키는 데 있다.

REFERENCES

1. Choi GY, Kim JM, Seo JH, Sohn HJ. Becoming a doctor through learning anatomy: narrative analysis of the educational experience. Korean J Phys Anthropol. 2009;22:213-24.
2. Cho KJ, Kim S. Changes of recognition to death before and after observation on the cadaver dissection to paramedical students. Korean J Phys Anthropol. 2018;31:159-65.
3. Ryu JH, Yu SK. A content analysis of self-reflection journals after cadaver dissection practice by dental students. Anat Biol Anthropol. 2025;38:113-22.
4. Kristjansson K. Aristotelian character education. Abingdon (Oxon): Routledge; 2015.
5. Beauchamp TL, Childress JF. Principles of biomedical ethics. 8th ed. New York (NY): Oxford University Press; 2019.
6. Pellegrino ED, Thomasma DC. The virtues in medical practice. New York (NY): Oxford University Press; 1993.
7. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ. 2004;38:314-26.
8. Roh H, Lee HJ, Park SB, Yang JH, Kim DJ, Kim SH, et al. Multiple mini-interview in selecting medical students. Korean J Med Educ. 2009;21:103-15.
9. Pau A, Jeevaratnam K, Chen YS, Fall AA, Khoo C, Nadarajah VD. The multiple mini-interview (MMI) for student selection in health professions training: a systematic review. Med Teach. 2013;35:1027-41.
10. Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB. Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: a BEME systematic review: BEME Guide No. 37. Med Teach. 2016;38:443-55.
11. Callwood A, Jeevaratnam K, Kotronoulas G, Schneider A, Lewis L, Nadarajah VD. Personal domains assessed in multiple mini interviews (MMIs) for healthcare student selection: a narrative synthesis systematic review. Nurse Educ Today. 2018;64:56-64.
12. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50:36-60.
13. Hamila S, Birchill K, Cao K, Hossain MN, Bullock S, Hodgson W, et al. AI-generated multiple mini interview (MMI) stations for medical school admissions: psychometric evaluation. JMIR Med Educ. 2026;12:e86208.
14. Menon T, Mahajan A, Powell D. Designing AI-resilient admissions interviews for health professions training in the age of generative AI. NPJ Digit Med. 2026;9:488.
15. Sharma R, Gordon M, Dharamsi S, Gibbs T. Systematic reviews in medical education: a practical approach: AMEE Guide 94. Med Teach. 2015;37:108-24.
16. Gordon M, Gibbs T. STORIES statement: publication standards for healthcare education evidence synthesis. BMC Med. 2014;12:143.
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10:39.
19. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368:l6890.
20. Lewin S, Booth A, Glenton C, Munthe-Kaas H, Rashidian A, Wainwright M, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implement Sci. 2018;13 Suppl 1:2.
21. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA. 2007;298:1002-9.
22. Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med. 2015;90:1067-76.
23. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.
24. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. Acad Med. 2004;79 10 Suppl:S40-2.
25. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ. 2009;43:767-75.
26. Eva KW, Reiter HI, Rosenfeld J, Trinh K, Wood TJ, Norman

www.kci.go.kr

GR. Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. JAMA. 2012;308:2233-40.
27. Reiter HI, Eva KW, Rosenfeld J, Norman GR. Multiple mini-interviews predict clerkship and licensing examination performance. Med Educ. 2007;41:378-84.
28. Kim DH, Hwang J, Kim EJ, Yoon HB, Shin JS, Lee S. How different are premedical freshmen who enter after introducing a multiple mini-interview in a medical school? Korean J Med Educ. 2014;26:87-98.
29. Kim JK, Kang SH, Lee HJ, Yang J. Can the multiple mini-interview predict academic achievement in medical school? Korean J Med Educ. 2014;26:223-9.
30. Lee HJ, Park SB, Park SC, Park WS, Ryu SW, Yang JH, et al. Multiple mini-interviews as a predictor of academic achievements during the first 2 years of medical school. BMC Res Notes. 2016;9:93.
31. Kim KJ, Nam KS, Kwon BS. The utility of multiple mini-interviews: experience of a medical school. Korean J Med Educ. 2017;29:7-14.
32. Kim KJ, Lee NY, Kwon BS. Benefits and feasibility of using videos to assess medical school applicants' empathetic abilities in multiple mini interviews. Med Sci Educ. 2021;31:175-81.
33. Griffin B, Wilson I. Associations between the Big Five personality factors and multiple mini-interviews. Adv Health Sci Educ Theory Pract. 2012;17:377-88.
34. Jerant A, Griffin E, Rainwater J, Henderson M, Sousa F, Bertakis KD, et al. Does applicant personality influence multiple mini-interview performance and medical school acceptance offers? Acad Med. 2012;87:1250-9.
35. McLaughlin JE, Singer D, Cox WC. Candidate evaluation using targeted construct assessment in the multiple mini-interview: a multifaceted Rasch model analysis. Teach Learn Med. 2017;29:68-74.
36. Breil SM, Forthmann B, Hertel-Waszak A, Ahrens H, Brouwer B, Schonefeld E, et al. Construct validity of multiple mini interviews-Investigating the role of stations, skills, and raters using Bayesian G-theory. Med Teach. 2020;42:164-71.
37. Renaud JS, Bourget M, St-Onge C, Eva KW, Tavares W, Salvador Loye A, et al. Effect of station format on the psychometric properties of multiple mini interviews. Med Educ. 2022;56:1042-50.
38. Ayub R, Yousuf N, Shabnam N, Ashraf MA, Afzal AS, Rauf A, et al. Investigating the internal structure of multiple mini interviews - a perspective from Pakistan. PLoS One. 2024;19:e0301365.
39. Szkwara JM, Bannatyne AJ, Asil M, Craig B, Stokes-Parish J, Bishop J. Leveling the playing field: evaluating measurement equivalence in MMIs between genders. Front Med (Lausanne). 2025;12:1639532.
40. Góes JS, de Paula SI, Daniel-Filho DA, Pires EMSG, Paes ÂT, Troster EJ. Effectiveness of multiple mini interviews in medical school admissions: assessment using the Big Five Personality framework. Einstein (Sao Paulo). 2025;23:eAO1352.
41. Duff RE, Katcher PA, Daniels RM, Ramaswamy V. The multiple mini interview as a dental school admission tool: can it predict noncognitive traits associated with professional behaviors? J Dent Educ. 2020;84:478-85.
42. Kok KYY, Chen L, Idris FI, Mumin NH, Ghani H, Zulkipli IN, et al. Conducting multiple mini-interviews in the midst of COVID-19 pandemic. Med Educ Online. 2021;26:1891610.
43. Carroll Turpin MA, Steele K, Matuk-Villazon O, Rowland K, Dayton CB, Horn KV. Rapid transition to a virtual multiple mini-interview admissions process: a new medical school's experience during the COVID-19 pandemic. Acad Med. 2021;96:1152-5.
44. Domes T, Bueckert S, Tetyurenko G, Hall D, Ironside A, Stobart K. Conducting a synchronous virtual multiple mini-interview using Webex for medical school admissions. Can Med Educ J. 2021;12:120-2.
45. Inzana KD, Vanderstichel R, Newman SJ. Virtual multiple mini-interviews for veterinary admissions. J Vet Med Educ. 2022;49:273-9.
46. Sabesan V, Kapur N, Zemanek K, Levitt D, Vu T, Van Erp A. Implementation and evaluation of virtual multiple mini interviews as a selection tool for entry into paediatric postgraduate training: a Queensland experience. Med Teach. 2022;44:87-94.
47. Hammond S, McLaughlin JE, Cox WC. Validity evidence for a virtual multiple mini interview at a pharmacy program. BMC Med Educ. 2023;23:551.
48. Callwood A, Harris J, Gillam L, Roberts S, Kubacki A, Christidis A, et al. Cross-sectional evaluation of an asynchronous multiple mini-interview (MMI) in selection to health professions training programmes with 10 principles for fairness built-in. BMJ Open. 2023;13:e074440.
49. Cork SC. A model for online delivery of multiple mini interviews. J Med Educ Curric Dev. 2023;10:23821205231183875.
50. Iramaneerat C, Maprapho P, Uakit O. Validity and reliability of the Multiple-Mini Interview (MMI) in medical student selection: a comparison of onsite and online platforms. Siriraj Med J. 2025;77:768-76.
51. Auton JC, Sturman D, Lekkas D, Turnbull D, Gray J, Canny BJ, et al. Virtual multiple mini interviews: implementation and evaluation for undergraduate medical, dental, and oral health degrees. Med Teach. 2025;47:1144-51.
52. Sa B, Sahu PK, Bodkyn C, Youssef F, Bethelmy A, Maharaj R,

www.kci.go.kr

et al. Piloting virtual multiple mini interviews (MMI) on undergraduate medical students: a multisource feedback analysis. Adv Med Educ Pract. 2025;16:761-72.
53. Yolanda S, Wisnu W, Wahjudi JM, Findyartini A. Adaptation of internet-based multiple mini-interviews in a limited-resource medical school during the coronavirus disease 2019 pandemic. Korean J Med Educ. 2020;32:281-9.
54. Kelly ME, Patterson F, O'Flynn S, Mulligan J, Murphy AW. A systematic review of stakeholder views of selection methods for medical schools admission. BMC Med Educ. 2018;18:139.
55. Roberts C, Khanna P, Rigby L, Bartle E, Llewellyn A, Gustavs J, et al. Utility of selection methods for specialist medical training: a BEME (best evidence medical education) systematic review: BEME Guide No. 45. Med Teach. 2018;40:3-19.
56. Yusoff MSB. Multiple mini interview as an admission tool in higher education: insights from a systematic review. J Taibah Univ Med Sci. 2019;14:203-40.
57. Lin JC, Lokhande A, Margo CE, Greenberg PB. Best practices for interviewing applicants for medical school admissions: a systematic review. Perspect Med Educ. 2022;11:239-46.
58. Wildermuth A. Predictive ability of multiple mini-interviews in admissions on programmatic academic achievement: a systematic review. J Allied Health. 2022;51:154-9.
59. Huang PH, Arianpoor A, Taylor S, Gonzales J, Shulruf B. Insights into undergraduate medical student selection tools: a systematic review and meta-analysis. J Educ Eval Health Prof. 2024;21:41.
60. Lin JC, Hu DJ, Scott IU, Greenberg PB. Evidence-based practices for interviewing graduate medical education applicants: a systematic review. J Grad Med Educ. 2024;16:151-65.

간추림 : 이 연구는 보건의료계열 학생 선발에서 다면인적성면접(MMI)이 무엇을 관찰할 수 있으며 그 결과를 어디까지 해석할 수 있는지 체계적으로 검토하였다. 2026년 7월 9일 6개 데이터베이스를 검색하여 952건을 확인했고, 최종 45편(1차 실증연구 33편, 체계적 문헌고찰 11편, 문항 구성 지침 1편)을 포함하였다. MMI 스테이션은 윤리적 딜레마, 대인관계에서의 진실성과 책임, 공감적 의사소통, 공적 책임의 네 범주로 정리되었다. 복수 스테이션과 평가자 구조는 전통적 면접보다 안정적인 방향의 근거를 보였으나 신뢰도는 설계에 따라 크게 달랐고, 입학 단계의 수행이 장기적인 임상윤리 행동을 예측한다는 근거는 충분하지 않았다. 비대면·생성형 AI 환경에서는 공정성, 구인 명료화, 인간의 검토와 책임이 추가 과제로 확인되었다. MMI는 인격 자체를 판정하는 도구라기보다 특정 상황에서 나타난 윤리적 판단과 관련 수행을 구조화해 표집하는 평가로 이해하는 것이 타당하다.

찾아보기 낱말 : 다면인적성면접, 입학전형, 의료윤리, 전문직업성, 의학교육

www.kci.go.kr