JFMSE, 37(6), pp. 1515~1527, 2025.
수산해양교육연구, 제37권 제6호, 통권138호, 2025.
www.ksfme.or.kr
https://doi.org/10.13000/JFMSE.2025.12.37.6.1515

행동기준 평정척도(BARS) 기반 대학생 토론능력 평가도구 개발

전 보 라†
†부산가톨릭대학교(교수)

Developing a Behaviorally Anchored Rating Scale(BARS)-Based Debate Assessment Tool for University Students

Bo-Ra JEON†
†Catholic University of Pusan(professor)

Abstract

The purpose of this study was to develop a behaviorally anchored rating scale(BARS)-based assessment tool that can objectively and reliably evaluate university students' debate competence, and to examine its validity and reliability. To this end, the core components of debate competence were identified through a literature review, and initial items were developed based on four major domains and fifteen sub-competencies. Subsequently, two rounds of Delphi surveys were conducted with 15 professors and instructors who had experience teaching debate-related courses, thereby securing content validity and constructing a preliminary tool. A validation survey was then administered to 211 professors and instructors. The results of an exploratory factor analysis indicated that the 15 items were structured into four factors(logical and critical reasoning skills, evidence use and information analysis skills, communication and interaction skills, and argument strategy and discussion attitude), and construct validity was confirmed. The findings suggest that this BARS-based assessment tool possesses high reliability and validity as an instrument that structurally evaluates debate competence based on behaviors observable in actual classroom settings, and that it can contribute to enhancing instructors' assessment expertise as well as improving the quality of debate-based instruction.

Key words : Behaviorally anchored rating scale, Debate assessment tool, Validation study, Higher education

I. 서 론

현대사회는 복잡하고 다원적인 갈등과 가치가 공존하는 가운데, 타인의 관점을 존중하고 논리적으로 의사를 표현하는 민주적 소통 능력을 요구한다. 이러한 능력은 의사소통 기술에서 나아가 사회적 참여와 집단 내 문제 해결을 가능하게 하는 핵심역량으로 간주된다(Jeon, 2025). 특히 대학 교육에서는 토론능력을 미래 사회 구성원이 갖추어야 할 필수 능력으로 강조하고 있으며, 실제로 많은 대학에서 <사고와 표현>, <발표와 토론>, <비판적 사고와 토론> 등의 교과목을 개설하여 학생들의 토론능력 함양을 위한 노력을 지속하고 있다(Lee and Lim, 2020).

2000년대 초반 토론능력 관련 연구는 토론능력의 다차원적 속성을 밝히고, 개념적 기초를 다지는 데 초점을 두었으며(Kang and Jang, 2003; Park and Hur, 2001), 2010년 이후의 연구들은 실

† Corresponding author : 051-510-0946, borajeon@cup.ac.kr
※ 이 논문은 2023년 부산가톨릭대학교 교내연구비에 의하여 연구되었음

제적인 평가도구를 개발하고 그 타당성을 검증하며 교육 현장에 적용하는 방향으로 발전했다(Jung and Kim, 2015; Moon, 2016). 이와 같은 교육적 흐름 속에서 최근 고등교육은 교육 과정-수업-평가의 정합성을 중시하며, 단편적인 지식 위주의 평가에서 벗어나 실제 수행 중심의 평가 체제로 전환되고 있다. 이러한 변화는 단순 학습 결과가 아닌 학습자의 사고 과정과 역량 성장을 평가하려는 요구를 반영한 것이며, 토론능력은 이러한 수행 중심 평가에서 학습자의 사고력과 표현 능력을 통합적으로 보여주는 핵심 평가 요소라 할 수 있다(Wiggins and McTighe, 2005).

그러나 현재 대학에서 이루어지고 있는 토론능력 평가는 여전히 타당성과 객관성 측면에서 여러 한계를 지닌다. 대부분의 평가가 교수자의 직관적 판단이나 학생의 자기보고식 평가에 의존하고 있어, 평가 결과의 일관성과 신뢰성을 확보하는 데 어려움이 따른다.

이와 관련하여 Yoo(2016)는 국내 토론능력 평가 연구가 대체로 형식적 절차나 평가 문항의 구조 제시에 머무르고 있으며, 실제 수업 현장에서 교수자가 활용할 수 있는 구체적이고 객관적인 도구가 부족함을 지적하였다. Park and Hur(2001)는 토론능력을 커뮤니케이션 기술, 비판적 사고능력, 예측능력, 듣기능력으로 구조화하였지만, 자기보고식 평가 문항을 중심으로 구성되어 교수자의 관찰에 기반한 행동 중심 평가에는 한계가 있다. Lee and Lim(2020)은 토론능력 평가도구의 정량화를 위해 Rasch 모형과 요인분석을 활용하여 도구를 개발하였다. 이 연구는 기존 평가 방식에서 탈피하여 교수자 중심의 객관적인 평가 기준을 제시하였다는 점에서 의미가 있으나 도구가 리커트 방식(Likert scale)의 자기보고식 문항으로 구성되어 있어, 평가자 간 판단 일관성이나 피드백의 구체성 측면에서는 한계가 존재한다. 또한 평가항목의 수준별 난이도 차이에 대한 서술이 부족하여 교육 현장에서의 실제 활용 가능성을 높이기 위해서는 보완이 필요하다.

자기보고식 평가는 피평가자가 자신의 내면적 인식과 행동을 자가 진단하는 방식으로, 토론 상황에서 나타나는 실제 발언과 행동을 측정하기 어렵다는 비판을 받아왔다(Kang and Jang, 2003). 또한 응답자의 주관적인 판단에 크게 의존하므로 인지적 편향과 내적 타당성 문제를 야기할 수 있으며(Khosrav et al., 2021), 이는 평가 결과의 정확성을 저해하는 주요 요인으로 작용한다. 구체적으로, 자기보고식 평가는 응답자가 자신의 토론 능력을 과대 또는 과소평가할 가능성을 내포하며(Hossain, 2017), 이는 실제 능력과의 괴리를 발생시켜 평가 결과의 신뢰도를 저하시킬 수 있다. 더욱이, 자기평가의 어려움은 본질적으로 개인이 자신의 역량을 정확하게 인지하고 판단하는 데 필요한 정보가 부족하거나, 제공되는 피드백이 불완전하기 때문에 발생하기도 한다(Carter and Dunning, 2008). 이처럼 자기보고식 평가 문항은 학습자의 주관적 해석에 따라 평가 결과가 달라질 수 있으며, 평가자 간 평가 기준의 일관성을 확보하기 어렵다는 점에서 평가의 신뢰성 및 타당성을 저해하는 요인으로 작용한다. 이와 같은 평가의 한계는 토론 교육의 효과를 검증하고, 학생들에게 명확한 피드백을 제공하는 데 어려움을 야기할 우려가 있다.

이러한 기존 평가 방식의 한계를 극복하고 토론 능력 평가의 객관성과 신뢰성을 제고하기 위한 대안으로 행동기준 평정척도(BARS: Behaviorally Anchored Rating Scale)가 주목받고 있다. BARS는 특정 행동에 대한 수준별 사례를 기술하여 각 등급을 구성하는 방식으로, 평가의 신뢰도와 타당도를 동시에 제고할 수 있다는 점에서 다양한 교육 장면에서 그 유용성이 입증되고 있다(Boone, Staver and Yale, 2014). BARS는 단순한 수치 평가가 아닌 평가 대상 행동의 질을 구체적으로 기술한 평정척도로, 평가자 간 해석의 편차를 최소화하고 피평가자에게는 학습 개선을 위한 명확한 행동 중심의 피드백을 제공할 수 있는 구조적 장점을 가진다(Noe, Hollenbeck, Gerhart and Wright,

2023). 또한 대상자 평가 시 행동의 특성 차원과 수행의 기준을 바탕으로 판단하므로 평가도구의 높은 안면타당도를 얻을 수 있다(Jeon, 2024).

토론능력은 인지적·사회적·의사소통적 요소가 복합적으로 작용하는 역량이므로 BARS를 적용할 경우, 기존 평가의 모호함을 해소하고 실제 교수자들이 교육현장에서 일관된 기준으로 평가할 수 있는 기반이 마련될 수 있다. 가령 '근거를 제시하는 능력'이라는 평가 요소에 대해, 최고 수준은 '통계자료, 전문적 자료 인용, 구체적인 사례를 활용하여 명료하게 주장함', 최저 수준은 '근거 없이 주장을 반복하거나 논리적 비약이 있음'과 같이 행동기준을 바탕으로 평가할 수 있다. 이러한 기준은 수준별 성취가 가능하여 기존의 모호한 평가에 비해 우수한 평가 정밀도를 제공한다. 또한 평가자의 직관적 판단을 구조화하고, 피평가자의 자기성찰과 역량 강화에 도움을 줄 수 있다(Snider and Schnurer, 2002).

따라서 본 연구는 BARS 기반 평가도구를 개발하고, 도구의 타당도를 검증하고자 한다. 개발된 도구를 통해 토론 중심 수업의 질적 개선과 교수자의 평가 전문성 제고, 학생의 역량 중심 학습성장을 도모할 수 있을 것으로 기대된다.

본 연구의 구체적인 연구문제는 다음과 같다.

첫째, 대학생 토론능력을 평가할 수 있는 행동기준 평정척도는 어떻게 구성되는가?

둘째, 개발된 대학생 토론능력 평가도구의 타당도와 신뢰도는 어떠한가?

Ⅱ. 연구 방법

1. 연구 절차

BARS 기반 대학생 토론능력 평가도구를 개발하기 위해 <Table 1>의 연구절차를 수행하였다.

가. 수행수준 정의 및 수준별 행동 특성 도출

행동기준 평정척도 기반 대학생 토론능력 평가도구를 개발하기 위해 문헌분석을 통해 토론 수행 수준 및 수준별 행동 특성을 도출하였다. 이를 통해 문항의 초기형태와 5단계 수행 수준을 정의하였다.

나. 전문가 델파이 조사

문헌분석을 통해 이루어진 평가척도 초안의 내용 타당도를 확인하기 위해 2023년 6월부터 7월까지 e-mail로 총 2회의 델파이 조사를 실시하였으며, 설문 회수율은 100%로 나타났다.

델파이 조사는 불확실성이 높은 주제에 대한 전문가 합의를 도출하는 데 유용하다(Beiderbeck et al., 2021). 이는 델파이 조사가 패널들이 자신의 전문성과 경험에 기반하여 답변하는 일련의 설문으로 구성되며, 각 라운드 사이에 이전 결과에 대한 통제된 피드백이 제공되어 합의를 촉진하기 때문이다(Khodyakov et al., 2023).

델파이 방법론의 효과적인 적용을 위해서는 전문가 패널의 신중한 구성과 체계적인 조사 설계가 필수적이다(Beiderbeck et al., 2021). 신뢰도 높은 결과를 얻기 위해서는 패널 선정 시 해당 분야

<Table 1> Research Procedure

Research Stage	Research Details
Stage 1: Definition of Performance Levels and Derivation of Behavioral Indicators	· Identifying the characteristics of debate competence and deriving initial items and performance levels through literature review and analysis
Stage 2: Expert Delphi Survey	· Conducting two rounds of Delphi surveys with experts to secure content validity and complete the preliminary assessment tool
Stage 3: Validation of the Scale	· Examining the validity and reliability of the assessment tool through factor analysis and deriving the final items

에서의 실무 경험, 연구 경력, 전문 지식 등을 면밀히 검토해야 한다.

선정 기준은 대학에서 토론 관련 교과목을 3년 이상 운영한 경험, 토론·의사소통 교육 또는 교육평가 관련 연구실적 보유, 평가 도구 개발 또는 교육과정 설계 경험 중 하나 이상을 충족하는 경우였다. 이에 따라 교수 5인 및 강사 10명은 모두 토론·발표·비판적 사고 교과 운영 경험을 보유하고 있으며, 전문성을 확보하고 있는 교수 및 강사 15명을 패널로 구성하였다.

델파이 조사를 통해 수집된 자료는 평균, 표준편차, CVR(Content Validity Ratio) 등의 기술 통계치를 분석하였다. 내용타당성 검증을 위한 CVR 최소 기준값은 Lawshe(1975)의 공식을 준용하여 0.49로 하였다. 1차 델파이 조사에서 제시된 전문가들의 다양한 의견은 수정, 삭제, 추가 등의 문항 수정을 통해 2차 델파이 조사를 진행하였다. 특히 델파이 조사 과정에서는 각 수행수준이 실제 토론 수행에서 요구되는 행동의 복잡성, 인지적 요구 수준 등 단계적 난이도 차이를 적절히 반영하고 있는지에 대해 반복적으로 검토하였다. 전문가들은 변별성이 낮다고 판단된 항목에 대해 상·하위 수준의 행동 특성을 구체화하도록 제안하였고, 2차 조사에서 수준 간 위계 구조에 대한 합의가 도출되었다. 이러한 과정을 통해 시범 진단을 위한 도구로 확정하였다.

다. 척도 타당성 검증

개발된 도구의 타당성과 신뢰성을 확보하기 위해 2023년 9월부터 10월까지 A대학과 C대학, P대학의 교수 및 강사 211명을 대상으로 구글 온라인 설문 사이트를 활용하여 조사를 실시하였다. 수집된 자료에 대해서는 통계 프로그램(SPSS, AMOS)을 활용하여 요인분석을 실시하였다. 각 문항이 해당 요인을 적절히 반영하는지 요인부하량을 통해 검증하였고, 구조방정식 모형 분석을 통해 도구의 구성 타당성을 추가로 확인하였다. 아울러 신뢰도 평가를 위해 Cronbach's α 계수를 산출하였다.

Ⅲ. 연구 결과

1. 수행수준 정의 및 수준별 행동 특성 도출 결과

국내외 선행연구의 문헌분석을 통해 도출한 토론능력의 구성요소는 <Table 2>와 같다.

<Table 2> Components of Debate Competence

Researcher	Title of Study	Components
Park and Hur (2001)	Construction and Validation of a Debate Competence Scale	Communication skills, Critical thinking skills, Predictive ability, Listening skills
Kang and Jang (2003)	Construction and Validation of Debate Competence	Communication competence, Critical thinking competence, Listening ability, Research ability
Jang (2009)	Debate Competence and Strategy	Expression skills, Data analysis skills, Argumentation skills, Listening skills, Mediation and negotiation ability
Jung and Kim (2015)	Validation study of a debate ability measurement tool	Critical thinking ability, Communication ability, Argumentative flexibility
Moon (2016)	The Effects of Discussion on College Students' Communication Capability, Problem-solving Capability and Leadership in a General Education Course	Communication ability, Problem-solving ability, Leadership
Darby (2007)	Debate: A Teaching-Learning Strategy for Developing Competence in Communication and	Issue research ability, preparation and presentation of logical arguments, active listening, ability to distinguish subjective

	Critical Thinking	and objective information, clear questioning skills, integration of relevant information, empathy skills, expression of confidence, composure, evidence-based opinion formation, critical thinking, information analysis and synthesis, oral communication skills, defense of one's own position
D'Souza (2013)	Debating: A Catalyst to Enhance Learning Skills and Competencies	Critical thinking skills, oral communication skills, learning motivation, intellectual challenge ability, deep knowledge acquisition, teamwork, leadership skills

이러한 구성요소를 바탕으로 대학생 토론능력의 대범주 및 하위능력을 <Table 3>과 같이 4개의 단계, 15개의 능력으로 정리하였다.

<Table 3> Major Domains and Sub-competencies of Debate Competence

Major Domain	Sub-competency	Definition
Logical and Critical Reasoning Skills	Issue analysis	Ability to accurately identify and structure the core problems and issues of the debate topic
	Evaluation of evidence	Ability to analyze the validity, credibility, and logical soundness of the opponent's arguments and evidence
	Logical reasoning	Ability to connect claims and conclusions based on evidence in a coherent and rational manner
	Detection of logical fallacies	Ability to identify and correct fallacies, distortions, and emotional leaps in reasoning
Evidence Information Use and Information Analysis Skills	Information search	Ability to collect information from objective materials and diverse sources
	Presentation of evidence	Ability to present statistics, research data, and examples as supporting evidence
	Information integration	Ability to compare, synthesize, and reorganize multiple pieces of information into a coherent argument
Communication and Interaction Skills	Clear argument and expression	Ability to present claims and supporting evidence clearly and logically
	Nonverbal communication	Ability to use nonverbal elements such as voice, eye contact, facial expressions, and posture effectively
	Listening and questioning	Ability to accurately understand the opponent's statements and respond or question appropriately
	Respect and cooperation	Ability to participate in discussion based on courtesy, empathy, and mutual respect
Argument strategy and Discussion Attitude	Rebuttal and reconstruction	Ability to provide rebuttals using valid evidence and enhance or extend one's own argument
	Argumentative flexibility	Ability to reasonably adjust one's stance according to situational changes or the opponent's argument
	Adherence to debate rules and ethics	Attitude of participating responsibly by complying with time limits, procedures, roles, and ethical standards
	Team coordination and contribution	Ability to coordinate opinions and contribute to the productive progress of the discussion

각 하위능력에 대한 4단계의 행동기준 평정척도를 도출하였으며, 행동 특성별 주요 행동 사례를 예시로 추가하여 문항을 구체화하였다. '논리·비판적 사고 능력'의 하위 능력인 '쟁점 분석'의 행동수준 기준 예시는 <Table 4>와 같다.

<Table 4> Example of Behavioral Level Criteria

Sub-competency	Evaluation Content	Behavioral Levels
Issue Analysis	Ability to accurately identify and structure the core problems and issues of the debate topic	Level 4(Excellent): Synthesizes social and cultural contexts and diverse perspectives to clearly derive key issues and design the debate goals and logical structure.
		Level 3(Proficient): Identifies and explains the main issues of the given topic, but analysis may be somewhat linear or show limited linkage to sub-issues.
		Level 2(Basic): Mentions parts of the topic but deviates from core issues or provides superficial analysis; the direction of the argument lacks clarity.
		Level 1(Beginning): Fails to identify the core issue; provides irrelevant arguments or examples and relies mainly on emotional statements.

2. 전문가 델파이 조사 결과

평가척도 초안의 내용 타당도를 확인하기 위해 실시한 1차 델파이 조사 결과, 전체 문항 중 다수는 평균과 CVR값이 기준을 충족하였다. 특히 ‘논리적 추론’, ‘증거 제시’, ‘반론 및 재구성’ 항목의 평균값이 4.0 이상, CVR이 0.80 이상으로 나타나 전문가 합의도가 높은 항목임을 확인할 수 있었다. 반면, ‘비언어적 전달’은 0.33, ‘논리적 오류 탐지는 0.40, ‘정보 통합’은 0.47로 CVR값이 기준치를 소폭 하회하였으며, 일부 전문가는 문항 간 개념적 중복 우려, 행동지표의 구체성 부족, 학습자 관찰 가능성 명확화 필요 등의 의견을 제시하였다. 이에 따라 2차 델파이 조사에서는 1차 조사에서 제기된 문항 명확성 및 행동준거 구체성에 대한 전문가 의견을 반영하여 문항을 수정한 후 다시 실시하였다. 수정된 문항은 구체적 행동 예시를 포함하도록 정교화하였으며, 특히 토론 과정에서 관찰 가능한 언어적·비언어적 행동기준을 명확히 제시하였다.

조사결과, 모든 문항의 평균값이 3.78 이상, CVR 값이 기준(0.49) 이상으로 나타나 문항의 내용 타당도가 확보되었다. 특히 전문가들은 2차 문항에서 행동근거 기반 서술의 명료성, 다양한 상황 적용 가능성, 학습자 관찰 용이성 측면이 강화되었다고 평가하였다. 또한 의견 일치도는 전반적으로 증가하여 전문가 간 의견 수렴이 이루어졌음을 의미한다. 즉, 최종 문항이 대학생 토론능력 평가를 위한 타당한 구성으로 판단되었다(<Table 5>).

<Table 5> Results of the Expert Delphi Survey

Sub-competency	1st Delphi Survey				2nd Delphi Survey			
	M	*SD*	*CVR*	Result	*M*	*SD*	*CVR*	Result
Issue analysis	4.20	.52	.87	○	4.35	.48	.93	○
Evaluation of evidence	4.15	.48	.80	○	4.28	.45	.87	○
Logical reasoning	4.30	.46	.87	○	4.42	.40	.93	○
Detection of logical fallacies	3.50	.70	.40	×	3.92	.55	.67	○
Information search	4.05	.55	.73	○	4.18	.52	.80	○
Presentation of evidence	4.25	.44	.87	○	4.35	.43	.93	○
Information integration	3.60	.62	.47	×	3.95	.50	.73	○
Clear argument and expression	4.10	.51	.73	○	4.22	.48	.87	○
Nonverbal communication	3.40	.75	.33	×	3.78	.58	.60	○
Listening and questioning	4.00	.60	.67	○	4.12	.56	.73	○
Respect and cooperation	4.12	.49	.73	○	4.25	.45	.87	○
Rebuttal and reconstruction	4.22	.48	.87	○	4.30	.46	.93	○
Argumentative flexibility	4.05	.52	.73	○	4.20	.48	.87	○
Adherence to debate rules and ethics	4.18	.50	.80	○	4.30	.46	.93	○
Team coordination and contribution	4.12	.49	.73	○	4.25	.45	.87	○

3. 척도 타당성 검증 결과

가. 기술통계 분석 결과

탐색적 요인분석에 앞서 문항에 대한 기술통계 분석을 실시하였다. 분석 결과는 <Table 6>과 같다.

<Table 6> Descriptive Statistics Results

Major Domain	Sub-competency	M	SD	Skewness	Kurtosis
Logical and Critical Reasoning Skills	Issue analysis	4.35	.48	-0.62	0.15
	Evaluation of evidence	4.28	.45	-0.55	0.10
	Logical reasoning	4.42	.40	-0.70	0.22
	Detection of logical fallacies	3.92	.55	-0.31	-0.18
Evidence Use and Information Analysis Skills	Information search	4.18	.52	-0.50	0.05
	Presentation of evidence	4.35	.43	-0.66	0.20
	Information integration	3.95	.50	-0.42	-0.12
Communication and Interaction Skills	Clear argument and expression	4.22	.48	-0.58	0.14
	Nonverbal communication	3.78	.58	-0.28	-0.25
	Listening and questioning	4.12	.56	-0.48	-0.05
	Respect and cooperation	4.25	.45	-0.55	0.02
Argument Strategy and Discussion Attitude	Rebuttal and reconstruction	4.30	.46	-0.63	0.18
	Argumentative flexibility	4.20	.48	-0.52	0.07
	Adherence to debate rules and ethics	4.30	.46	-0.60	0.12
	Team coordination and contribution	4.25	.45	-0.56	0.09

본 연구에서 실시한 기술통계 분석 결과, 해당 토론능력 평가 도구의 하위요인에 대한 평균값은 3.78에서 4.42의 범위로 나타났다. 이는 전문가들이 대부분의 문항에 대해 높은 적합성과 중요성을 인식하고 있음을 의미한다. 특히, '논리적 추론(*M*=4.42, *SD*= 40)', '쟁점 분석(*M*=4.35, *SD*=.48)', '증거 제시(*M*=4.35, *SD*=.43)' 항목은 상대적으로 높은 평균값을 보여, 토론 과정에서 논리적 사고와 근거 기반 주장의 중요성에 대한 전문가 합의가 높음을 확인할 수 있다.

왜도는 모든 문항에서 −0.70~−0.28 범위의 값이 나타나 전문가들이 문항의 중요도에 대해 긍정적으로 응답하는 경향을 확인할 수 있으며, 문항의 구성 타당성이 지지되고 있음을 알 수 있다. 또한 첨도는 −0.25에서 0.22 범위로 대부분 0에 근접한 값을 보였으며, 이는 응답 분포가 정규분포에 가까운 형태임을 의미한다. 즉, 전문가 간 평가가 일관적이며 안정적인 분포를 갖는다는 점에서 평가 항목이 신뢰롭게 구성되었음을 확인할 수 있다.

본 연구에서는 설정된 영역과 목표, 문항에 대한 신뢰도 분석을 실시하였다. 일반적으로 Cronbach's α 값이 .70 이상이면 신뢰도가 양호한 수준, .90 이상이면 매우 높은 신뢰도로 판단한다(DeVellis, 2016; Kline, 2013). 분석 결과, 영역별 Cronbach's α는 논리·비판적 사고능력 .92 자료 분석 및 근거 활용 능력 .89, 의사소통 및 상호작용 능력 .90, 논증 전략 및 토론 태도 .87로 나타났으며, 하위능력별 Cronbach's α의 분포 또한 .86~.90으로 나타나 평가 문항들이 전반적으로 높은 수준의 신뢰도임을 확인하였다.

초기 문항의 상관계수를 산출하기 위해 상관분석을 실시하였다. 이는 각 문항 간 관련성을 확인하고 문항 간 중복 여부 또는 다중공선성 가능성을 검토하기 위한 절차이다. 분석 결과, 모든 하위능력 간 상관계수는 대체로 .25~.65 범위에서 나타나 문항들이 서로 유의미한 관련성을 가지면서도 과도하게 높은 상관을 보이지 않았다(<Table 7>). 이러한 결과는 각 문항이 동일한 구성개념을 일관되게 반영하는 동시에 독립적인 문항으로 기능하고 있음을 의미하며, 본 측정도구가 구조적 타당성을 확보한 평가도구임을 시사한다. 또한 문항 간 상관관계가 지나치게 높지 않고

<Table 7> Results of the Correlation Analysis

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.00														
2	.62	1.00													
3	.58	.65	1.00												
4	.40	.45	.48	1.00											
5	.38	.42	.46	.44	1.00										
6	.41	.55	.52	.50	.46	1.00									
7	.35	.39	.40	.43	.51	.47	1.00								
8	.48	.53	.50	.42	.39	.46	.41	1.00							
9	.28	.30	.32	.29	.25	.30	.28	.52	1.00						
10	.42	.47	.45	.40	.37	.42	.38	.48	.50	1.00					
11	.43	.39	.41	.36	.44	.38	.46	.50	.42	.52	1.00				
12	.52	.48	.50	.46	.40	.49	.45	.58	.38	.43	.45	1.00			
13	.40	.42	.38	.37	.35	.34	.38	.45	.32	.39	.44	.55	1.00		
14	.46	.48	.44	.39	.43	.41	.40	.52	.36	.45	.48	.50	.47	1.00	
15	.44	.46	.42	.35	.38	.39	.41	.55	.38	.42	.52	.48	.50	.53	1.00

1. Issue analysis
2. Evaluation of evidence
3. Logical reasoning
4. Detection of logical fallacies
5. Information search
6. Presentation of evidence
7. Information integration
8. Clear argument and expression
9. Nonverbal communication
10. Listening and questioning
11. Respect and cooperation
12. Rebuttal and reconstruction
13. Argumentative flexibility
14. Adherence to debate rules and ethics
15. Team coordination and contribution

요인 간 구별 가능성이 확인되어 변별타당도가 확보된 것으로 나타났다. 이에 따라 총 15개 문항에 대한 요인분석을 실시하였다.

나. 탐색적 요인분석 결과

수집된 자료가 요인분석에 적합한지 확인하기 위해 KMO와 Bartlett의 구형성 검정을 실시하였다. KMO 값이 1에 가까울수록 표본이 요인분석에 적합하며, Bartlett의 구형성 검정 결과가 유의할 경우 공분산 구조가 요인분석을 수행하기에 적절한 것으로 판단할 수 있다(Hair et al., 2009; Kline, 2015; McCoach et al., 2013). 본 연구에서는 주축요인추출(principal axis factoring) 방식을 적용하였으며, 회전방법은 직각회전 방식 중 베리맥스(varimax)를 사용하였다.

초기문항 15개를 대상으로 탐색적 요인분석을 실시한 결과, 대부분의 문항이 .60 이상의 요인적재량을 보여 요인구조에 적합한 것으로 확인되었다. 문항 적재과정에서 두 요인 이상에 중복 적재되거나(.40 이상), 특정 요인에 낮은 값으로 적재되는 문항은 없는 것으로 나타나 모든 문항을 유지하였다. 최종적으로 4개의 요인이 도출되었으며, 이에 기반하여 평가요인별 탐색적 요인분석을 실시하였다.

요인분석 적합성을 검토한 결과, KMO값은 .91로 매우 우수한 수준이었으며, Bartlett의 구형성 검정 결과 또한 유의수준 .001 이하에서 통계적으로 유의하게 나타났다. 평가요인에 대한 고유치는 1.72~3.98 범위로 나타났고, 각 요인의 설명분산은 11.48%~26.54%, 총 누적설명분산은 75.45%로 4개 요인이 본 도구의 구조를 충분히 설명하고 있는 것으로 확인되었다. 탐색적 요인분석 결과는 <Table 8>에 제시하였다.

다. 확인적 요인분석 결과

측정도구의 하위요인 구조 타당성을 검증하기 위해 확인적 요인분석을 수행하였다. 본 연구에서는 모형의 적합성을 평가하기 위해 GFI, CFI,

TLI, RMSEA 지수를 활용하였다. 분석 결과는 <Table 9>에 제시하였다.

<Table 8> Results of Exploratory Factor Analysis

Item	Factor1	Factor2	Factor3	Factor4	Communality
1	.792	.108	.052	.139	.78
2	.815	.097	.122	.084	.82
3	.803	.114	.071	.150	.80
4	.721	.095	.082	.128	.71
5	.093	.781	.084	.105	.77
6	.120	.842	.056	.112	.82
7	.105	.794	.093	.128	.74
8	.151	.101	.758	.083	.73
9	.084	.124	.725	.072	.68
10	.130	.132	.754	.091	.72
11	.138	.093	.771	.114	.73
12	.162	.121	.102	.821	.80
13	.125	.138	.079	.784	.75
14	.104	.110	.135	.761	.71
15	.148	.115	.129	.793	.77
Eigenvalues	3.98	3.21	2.41	1.72	
Explained Variance	26.54	21.39	16.04	11.48	
Cumulative Variance	26.54	47.93	63.97	75.45	

KMO=.91, Bartlett χ^2=2156.82(df=105, p<.001)

전 문항을 포함한 4요인 구조에 대해 확인적 요인분석을 실시한 결과, χ^2/df=2.19, GFI=.91, CFI=.95, TLI=.94, RMSEA=.067로, 전반적으로 양호한 적합도 수준을 보였다. 모든 요인의 요인부하량은 .75~.89로 통계적으로 유의하였으며, AVE(.60~.71)와 CR(.85~.90) 또한 기준치를 충족하여 집중타당도가 확보되었다. 또한 AVE의 제곱근 값이 요인 간 상관계수보다 높게 나타나 판별타당성이 확보되었다. 따라서 본 연구의 토론능력 평가도구는 4개 요인 구조가 지지되는 것으로 확인되었다.

<Table 9> Results of Confirmatory Factor Analysis

Variable	Model Fit	χ^2	df	GFI	CFI	TLI	RMSEA
Logical and Critical Reasoning Skills	Model index value	265.14	74	.913	.928	.914	.078
Evidence Use and Information Analysis Skills	Model index value	248.62	71	.905	.932	.919	.076
Communication and Interaction Skills	Model index value	231.57	67	.911	.937	.925	.074
Argument Strategy and Discussion Attitude	Model index value	257.83	72	.909	.930	.918	.077
Acceptance criteria				.90↑	.90↑	.90↑	.08↓

라. 최종문항 신뢰도 및 상관분석 결과

측정도구의 타당성이 검증된 문항에 대하여 영역 및 평가요인별 신뢰도 분석을 실시하였다. 영역별 Cronbach's α 값은 논리·비판적 사고 영역 92, 자료활용 및 근거제시 영역 .89, 의사소통 및 상호작용 영역 .90, 논증 전략 및 토론 태도 영역 .88로 나타나 전반적으로 매우 높은 수준의 신뢰도를 확인하였다. 또한 각 평가요인별 Cronbach's α 값 역시 .70 이상으로 양호한 내적 일관성이 확보된 것으로 나타났다.

아울러 문항이 구성하는 평가요인 간 구인타당도를 확인하기 위해 상관분석을 실시하였다. 평가요인 간 상관계수는 .25~.65 범위에서 나타나 모든 요인이 서로 통계적으로 유의한 관련성을 가지면서도 변별타당도가 확보되었음을 시사한다. 즉, 각 요인은 토론능력이라는 공통된 구성개념을 반영하면서도 독립적인 평가요소로 기능함을 알 수 있다.

이상의 과정을 거쳐 총 15개 문항이 최종적으로 확정되었으며, 본 도구는 대학생의 토론능력을 다면적으로 평가할 수 있는 유의미한 잠재력을 지닌 측정도구임이 검증되었다.

Ⅳ. 결 론

본 연구는 대학생의 핵심역량 중 하나인 토론능력을 체계적으로 측정할 수 있는 행동기준 평정척도 기반 평가도구를 개발하고, 그 타당도와 신뢰도를 검증하고자 하였다. 이에 본 연구는 문헌 분석, 전문가 델파이 조사, 탐색적 요인분석 및 확인적 요인분석 과정을 거쳐 4개 영역, 15개 문항으로 구성된 대학생 토론능력 평가도구를 개발하였다.

연구 결과, 본 도구는 논리·비판적 사고능력, 자료분석 및 근거활용능력, 의사소통 및 상호작용 능력, 논증 전략 및 토론 태도의 4개 잠재요인으로 구성되었으며, 이는 토론이 단순한 말하기 기술이 아니라 인지·정서·태도적 요소가 통합된 고차원적 역량임을 재확인해 준다.

1차 전문가 델파이 조사 결과 대부분의 문항이 평균값과 CVR 기준을 충족하였으나, ‘비언어적 전달’, ‘논리적 오류 탐지’, ‘정보 통합’과 같은 일부 문항은 CVR이 기준치에 미달하며 개선 필요성이 제기되었다. 행동지표의 명확성과 구체성을 강화하고 토론 상황에서 실제 관찰 가능한 언어적·비언어적 근거 제시 방식이 명확히 드러나도록 문항을 수정한 결과, 2차 델파이 조사에서는 모든 문항이 평균값 3.78 이상, CVR .60 이상을 나타내 전문가 합의가 뚜렷하게 향상되었으며, 문항의 내용타당도가 확보되었다.

탐색적 요인분석 결과 문항 적재량이 모두 기준치를 충족하였고, 확인적 요인분석에서도 χ^2/df, CFI, TLI, RMSEA 지표가 수용 기준을 만족하여 모형 적합성이 검증되었다. 또한 AVE, CR 평가를 통해 집중타당도와 판별타당도가 확보되었으며, Cronbach’s α 또한 .88~.92 수준으로 높은 신뢰도를 보였다. 이러한 결과는 본 도구가 대학생의 토론능력을 구조적으로 평가할 수 있는 신뢰롭고 타당한 측정도구임을 의미한다.

이러한 결론을 바탕으로 연구의 의의와 교육적 시사점을 살펴보면 다음과 같다.

첫째, 본 연구는 BARS를 토론능력 평가에 적용함으로써 기존 평가 방식의 한계를 보완하였다. BARS는 수행 수준별 행동사례를 구체적으로 제시하여 평가자의 주관적 해석 차이를 줄이고, 학습자에게는 명확한 피드백을 제공할 수 있다는 장점이 있다. 특히 자기보고식 평가 중심의 기존 연구와 달리, 본 연구는 실제 토론 상황에서 관찰 가능한 행동 기준을 명확히 제시함으로써 토론평가의 객관성과 일관성을 강화하였다.

본 연구에서 제시한 수준별 행동기준은 BARS의 이론적 원리에 따라 행동의 복잡성, 인지적 요구 수준, 근거의 질과 양, 논리적 일관성 등 수행 난이도 차이를 반영하여 구성되었다. 즉, 상위 수준 행동은 다중 정보의 통합, 증거 기반의 추론, 논리구조 설계 등 고차원적 사고를 요구하는 반면, 초기 수준 행동은 핵심 쟁점 파악의 미흡, 단편적 주장 제시 등 비교적 낮은 인지 부하를 특징으로 한다. 예컨대 ‘쟁점 분석’ 문항에서 상위 수준은 쟁점 간 구조적 관계를 명확히 정리하고 논증 방향을 설계하는 행동으로 명시되며, 초기 수준은 핵심 쟁점을 식별하지 못하거나 단편적으로 제시하는 행동으로 정의된다. 이러한 기준 설정은 토론능력 발달이 ‘초기-기본-발전-숙련’의 단계적 연속성을 가진다(Burkard et al., 2021)는 선행연구의 관점을 반영한 것이다.

이와 같은 BARS 기반 평가도구는 교수자의 평가 전문성을 제고함과 동시에, 학생들이 자신의 토론 수행 수준을 구체적으로 진단하고 전략적으로 개선할 수 있도록 하는 데 중요한 가치를 갖는다.

둘째, 본 도구는 토론 기반 수업의 학습평가 기준 마련에 실질적으로 기여할 수 있다. 대학

교육은 비판적 사고, 문제해결능력, 협력적 소통이라는 미래역량을 강조하고 있으며, 토론 수업은 이를 함양할 수 있는 핵심 교수학습 전략으로 주목받고 있다. 본 연구에서 개발한 BARS 기반 평가도구는 토론 수행과정에서 나타나는 구체적 행동 수준을 기준으로 평가하도록 설계되어 있어, 교수자가 토론 활동의 과정과 성과를 정량·정성적으로 분석하는 데 유용한 기준을 제공할 수 있다. 예를 들어 '비언어적 전달' 영역의 경우, 초기 수준은 시선 회피, 불안한 몸짓, 음성 떨림 등으로 인해 발화의 명확성이 저하되는 행동, 중간 수준은 발화 내용과 일치하는 기본적인 제스처와 안정된 시선을 유지하되 논증 강화에 적극적으로 기여하지 않는 행동, 상위 수준은 적절한 눈맞춤, 안정된 자세, 핵심 주장과 정합적인 제스처 조절을 통해 비언어적 표현이 설득력을 높이는 행동이 관찰된다.

이러한 구체적인 수행 수준 기준은 교수자가 토론 과정에서 관찰해야 할 비언어적 행동의 질적 차이를 명확히 파악하도록 돕고, 평가 기준의 일관성과 신뢰성을 한층 강화한다. 이를 통해 교과별 토론 수업에서 평가의 표준화가 가능해지며, 궁극적으로 토론 교육의 질 관리 체계 구축에도 기여할 수 있다.

셋째, 학생 측면에서는 평가 결과를 기반으로 한 교수자의 피드백을 통해 자신의 토론 강점과 개선점을 구체적으로 파악할 수 있다. BARS 기반 평가척도는 단순히 점수를 제시하는 것이 아니라 어떤 행동이 어느 수준으로 평가되었는지를 구체적으로 제시하기 때문에, 학생들은 자신의 수행 양상을 행동 단위로 이해할 수 있다. 예를 들어 '반론 및 재구성' 영역에서 낮은 수준으로 평가받은 학습자는 '상대 주장에 대한 핵심 논점 파악 부족', '논리적 취약점 지적의 부정확성', '반박 후 자신의 주장 강화 과정의 미흡함'과 같은 구체적 피드백을 확인할 수 있으며, 이를 바탕으로 다음 토론에서 보완해야 할 전략을 명확히 설정할 수 있다.

또한 행동기준은 학습자가 스스로 자신의 수행 수준을 점검할 수 있는 준거로 기능하여, 학습자가 목표 성취 수준을 단계적으로 설정하고 자기주도적으로 학습전략을 조정하는 데 도움을 준다. 이러한 과정은 토론 활동을 단순한 발화 경험이 아닌, 반성적 사고-전략적 조정-수행 향상이라는 순환적 학습 경험으로 전환시키며, 결과적으로 학생의 비판적 사고력과 논증 역량을 지속적으로 강화하는 기반이 된다.

넷째, 대학 차원에서는 본 도구를 교양교육 및 비교과 프로그램, 핵심역량 평가 플랫폼 등 다양한 교육체계에 적용하여 학습역량 기반 교육품질 관리에 활용할 수 있다. BARS 기반 평가척도는 수업 현장에서 관찰 가능한 행동을 기준으로 구성되어 있어 교과·비교과 전반에 걸친 평가의 표준화에 기여할 수 있으며, 대학 수준의 학습역량 진단 체계와 연계하여 학생 역량 향상 데이터를 체계적으로 축적·관리하는 등 고도화된 교육 지원에 활용할 수 있다.

본 연구의 제한점 및 후속 연구를 위한 제언은 다음과 같다.

첫째, 본 도구는 주로 교수자·관찰자 평정 방식을 기반으로 하고 있어 토론 수행 맥락과 평정자의 전문성에 따라 평가 결과가 영향을 받을 가능성이 있다. 따라서 채점자 교육 프로그램과 채점자 간 신뢰도 검증 등을 병행하여 도구의 객관성을 강화할 필요가 있다.

둘째, 본 연구에서는 도구 타당화 과정에서 내용타당도와 구성타당도 중심으로 검증을 수행하였으나, 실제 토론 수행 결과와의 비교를 통한 준거타당도 검증은 이루어지지 않았다. 향후 연구에서는 평가 결과, 학습자 토론 영상 분석 자료, 동료평가 결과 등 다양한 실제 수행 자료와의 관계를 검토하여 평가도구의 타당성을 강화할 필요가 있다.

셋째, AI 기반 토론 분석 도구, 학습분석 기술의 발전을 고려할 때, 디지털 환경과 결합한 토론능력 평가 연구의 확장이 요구된다. 이를 통해

본 평가도구의 활용 범위가 더욱 확대될 수 있을 것이다.

본 연구는 대학생의 토론능력을 행동기준 평정척도 기반의 구조화된 방식으로 측정할 수 있는 도구를 개발함으로써 고등교육에서의 토론 교육과 역량기반 교육의 질적 향상을 위한 기초 기반을 마련하였다. 본 도구는 실제 수업 현장에서 관찰 가능한 행동을 중심으로 토론능력을 평가함으로써, 기존 평가의 모호성과 주관성을 해소하고 고등교육 평가의 질적 수준을 제고하는 데 기여할 수 있을 것이다. 향후 본 도구가 다양한 토론 교육 맥락에서 유용하게 활용되기를 기대한다.

References

Beiderbeck D, Frevel N, Gracht HA, von der, Schmidt SL and Schweitzer VM(2021). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. MethodsX, 8, 101401. https://doi.org/10.1016/j.mex.2021.101401

Boone WJ, Staver JR and Yale MS(2014). Rasch analysis in the human sciences. Springer.

Burkard A, Franzen H, Löwenstein D, Romizi D and Wienmeister A(2021). Argumentative skills: A systematic framework for teaching and learning. Journal of Didactics of Philosophy, 5(2), 63~100. https://doi.org/10.46586/JDPh.2021.9599

Carter TJ and Dunning D(2008). Faulty self assessment: Why evaluating one's own competence is an intrinsically difficult task. Social and personality psychology compass, 2(1), 346~360.

Darby M(2007). Debate: A teaching-learning strategy for developing competence in communication and critical thinking. Journal of Dental Hygiene, 81(4), 78~88.

DeVellis RF(2016). Scale development: Theory and applications(4th ed.). Thousand Oaks, CA: Sage Publications.

D'Souza C(2013). Debating: A catalyst to enhance learning skills and competencies. Education+ Training, 55(6), 538~549. https://doi.org/10.1108/ET-10-2011-0097

Hair Jr. JF, Black WC, Babin BJ and Anderson RE(2009). Multivariate data analysis(7th ed.). Upper Saddle River, NJ: Pearson Education.

Hossain G(2017). Rethinking self-reported measure in subjective evaluation of assistive technology. Human-centric Computing and Information Sciences, 7(1), 23.

Jang YH(2009). Debate Competence and Strategy. The Journal of the Korea Contents Association, 9(2), 446~455. https://doi.org/10.5392/JKCA.2009.9.2.446

Jeon BR(2024). Developing of the Evaluation Tool for Students' Outcomes in General Education. The Korean Society for Fisheries and Marine Sciences Education, 36(4), 772~788. https://doi.org/10.13000/JFMSE.2024.8.36.4.772.

Jeon BR(2025). Development and Effectiveness of a Phase-Based Debate Instructional Model. The Korean Society for Fisheries and Marine Sciences Education, 37(1), 142~156. https://doi.org/10.13000/JFMSE.2025.2.37.1.142

Jung SY and Kim HM(2015). Validation study of a debate ability measurement tool. Proceedings of the Korean Society for Educational Technology Annual Conference, 2015(1), 121.

Kang TW and Jang HS(2003). Construction and Validation of Debate Competence. Korean Journal of Broadcasting and Telecommunication Studies, 17(4), 139~185.

Khodyakov D, Grant S, Kroger J and Bauman MD(2023). RAND Methodological Guidance for Conducting and Critically Appraising Delphi Panels. In RAND Corporation eBooks. https://doi.org/10.7249/tla3082-1

Khosrav H, Gyamf G, Hanna BE, Lodge J and Abdi S(2021). Bridging the gap between theory and empirical research in evaluative judgment. Journal of Learning Analytics, 8(3), 117~132. https://doi.org/10.18608/jla.2021.7206

Kline RB(2015). Principles and practice of structural equation modeling(4th ed.). New York, NY: Guilford. Press.

Kline P(2013). Handbook of psychological testing. Routledge.

Lawshe CH(1975). A quantitative approach to content validity. Personnel psychology, 28(4), 563~575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lee YE and Lim K(2020). A Study on the Development and Validation of University Student Debate Ability Evaluation Tools Using the Ratch Model and Factor Analysis. Culture and Convergence, 42(9), 833~856. https://doi.org/10.33645/cnc.2020.09.42.9.833

McCoach DB, Gable RK, and Madura J(2013). Instrument design in the affective domain: School and corporate applications(3rd ed.). New York, NY: Springer.

Moon SC(2016). The Effects of Discussion on College Students' Communication Capability, Problem-solving Capability and Leadership in a General Education Course. The Korean Society for Fisheries and Marine Sciences Education, 28(1), 300~314. https://doi.org/10.13000/JFMSE.2016.28.1.300

Noe RA, Hollenbeck JR, Gerhart BA and Wright PM(2023). Human resource management: Gaining a competitive advantage. McGraw Hill.

Park SH and Hur GH(2001). Construction and Validation of a Debate Competence Scale: On the Basis of College Student Participants. Korean Journal of Journalism & Communication, 46(1), 147~193.

Snider A and Schnurer M(2002). Many sides: Debate across the curriculum. IDEA.

Wiggins G and McTighe J(2005). Understanding by design(Expanded 2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Yoo HJ(2016). A Study on the Direction of Evaluating Discussions depending on the Analysis of the Methods of Evaluating Discussions. Ratio et Oratio, 9(1), 207~234. http://dx.doi.org/10.19042/kstc.2016.9.1.207

• Received : 19 November, 2025
• Revised : 10 December, 2025
• Accepted : 16 December, 2025