

생성형 AI를 활용한 음악 교과 서·논술형 평가 방안: 문항 개발과 채점에서 ChatGPT 활용 가능성을 중심으로

Exploring AI-Supported Restricted and Extended Essay Assessment in Music Education: The Potential of ChatGPT for Item Development and Scoring

유하양*

Hayang You 

초록 본 연구는 생성형 AI를 활용하여 음악 교과 서·논술형 평가 방안을 탐색하는 데에 목적이 있다. 이를 위해 대표적인 생성형 AI로 알려진 챗지피티(ChatGPT)를 통해 중학교 1학년 음악 수업에 적용할 수 있는 서·논술형 평가 문항을 개발하여 실제 평가를 실행한 후 채점하였다. 연구의 주요 결과는 다음과 같다. 첫째, 문항 개발 단계에서는 동일한 프롬프트를 제시할 때마다 ChatGPT가 생성한 문항의 내용이나 범위가 달라지는 것으로 나타났다. 또한 상위 버전 모델일수록 음악 용어에 대해 부연 설명을 덧붙이거나 문항을 구조화하여 제시하는 경향을 보였다. 둘째, 채점 단계에서는 다수의 자료를 한꺼번에 제시하는 프롬프트 방식보다 맞춤형 GPT를 활용했을 때 채점의 효율성과 일관성이 향상되는 것으로 나타났다. 특히 교사 채점 결과를 추가했을 경우 ChatGPT와 교사 채점 간 일치도가 유의미하게 높아졌다. 이를 통해 음악 교사의 평가 업무를 보조하는 도구로서 생성형 AI의 가능성과 한계를 확인하였으며, 생성형 AI의 효과적인 활용을 위해서는 교사의 비판적 검토와 전문적 판단이 필요함을 제안하였다.

주제어: 생성형 AI, 음악 교과 평가, 서·논술형 평가, 문항 개발, 채점, ChatGPT

Abstract This study investigates the potential of generative artificial intelligence (AI), particularly ChatGPT, as a supportive tool for item development and scoring of restricted and extended essay assessments in music education. For this purpose, assessment items appropriate for first-grade middle school music classes were developed using ChatGPT, then implemented and scored in an actual classroom setting. The findings indicate that during the item development phase, the generated items varied in content and scope despite using identical prompts. More advanced versions tended to offer clearer explanations of musical concepts and presented them in a more systematically organized manner. In the scoring phase, using customized GPT improved both efficiency and consistency compared to prompt-based approaches, which required the presentation of multiple materials simultaneously. Notably, incorporating teacher scoring results into the process significantly increased the level of agreement between ChatGPT and teacher scores. These findings highlight both the potential and limitations of generative AI as a tool to support music teachers' assessment practices. They also suggest that teachers' critical review and professional judgment are essential for the effective and responsible integration of generative AI in educational assessment.

Key words: generative AI, music assessment, essay assessment, item development, scoring, ChatGPT

* Corresponding author, E-mail: campanella@snu.ac.kr

Visiting Researcher, Seoul National University, Education Research Institute, 1 Gwanak-ro, Gwanak-gu, Seoul, South Korea

Received: 31 August 2025, Reviewed (Revised): 20 October (25 October) 2025, Accepted: 27 January 2026

© 2026 Korean Music Education Society.

I. 서론

급속도로 변화하는 미래 사회에 능동적으로 대응하기 위해서는 단순히 지식을 ‘암기’하고 ‘재생’하는 것이 아니라 이를 ‘분석’하고 ‘적용’하여 문제를 해결할 수 있는 역량이 요구된다. 이에 따라 학생 평가에 있어 복합적이고 고차원적인 사고 능력을 평가할 수 있는 여러 가지 방안들이 제안되었다. 이들 가운데 서·논술형 평가는 종래의 선택형 평가가 지니는 한계를 극복할 수 있는 대안이 될 수 있다는 점에서 정책적 중요성이 강조되어 왔다. 2022 개정 교육과정 총론에서는 서술형과 논술형 평가의 비중을 확대할 것을 강조하였으며, 2022 개정 음악과 교육과정에서는 다양한 평가 방법 가운데 하나로 서·논술형 평가를 제시하였다. 일부 시도교육청에서는 학기 단위 성적의 일정 부분 이상을 서·논술형 평가로 실시해야 한다는 조항을 「학업성적관리지침」에 포함하였다. 근래에는 대학수학능력시험에 서·논술형 문항의 도입 가능성이 제기되고 있다(Ministry of Education, 2023).

서·논술형 평가는 서술형이나 논술형 문항을 활용하여 학생이 직접 구성하거나 기술한 내용에 대해 전문적으로 판단하는 것으로, 분석력, 비판력, 종합력, 문제해결력, 창의력 등과 같은 고등 사고 능력을 평가할 수 있다는 점에서 교육적 가치를 지닌다(Back, 2002; Jeong & Back, 2025; McMillan, 2013). 이러한 평가 방식은 학생들의 지적 호기심과 탐구심을 자극하여 관련 정보를 해석하고 활용하는 과정에서 능동적이고 창의적인 사고력을 키울 수 있다(Downing, 2006). 또한 서·논술형 평가는 학습 결과뿐 아니라 과정에 대한 평가가 가능하다는 점에서 ‘과정중심평가’의 일환으로 학교 현장에서 유용하게 활용될 수 있다(Song et al., 2024). 이는 ‘학습을 위한 평가(assessment for learning)’ 혹은 ‘학습으로서의 평가(assessment as learning)’의 본질에 맞게 학생들의 학습 상태를 진단하고 결과에 대한 유의미한 피드백을 제공함으로써 평가 본래의 기능에 부합한다는 점에서 의미가 있다(Kim, 2020).

음악 교과에서 서·논술형 평가는 학습한 개념이나 원리에 대한 이해뿐만 아니라 어떠한 방식으로 사고하여 문제를 해결해 가는지를 파악함으로써, 지식과 기능 더 나아가 태도와 같은 정의적 영역에 대한 질적 평가가 가능하다(Lee, 2015). 이를 통해 음악적 성장과 발달에 대한 교육적 정보를 수집하여 학생에 대한 이해를 높일 수 있으며, 음악과 관련된 다양한 사회문제나 쟁점에 대한 자료를 종합적으로 탐색하여 자기의 생각이나 의견을 논리적으로 전개해 가는 탐구 중심 수업을 이끌 수 있다. 더불어 실기 평가를 비롯하여 자기평가, 동료평가, 보고서 평가, 포트폴리오 평가, 프로젝트 평가 등의 다양한 평가 방법과 병행함으로써 음악에 대한 심층적 이해 정도나 실기 능력의 향상 과정을 평가하는 데에 유용하다는 장점이 있다.

그러나 이러한 장점에도 불구하고, 학교 현장에서 교사들은 서·논술형 평가를 실행하는 데에 있어 여러 어려움을 겪고 있는 것으로 알려져 있다(Park et al., 2019; Song, 2022). 우선 서·논술형 평가 문항을 개발하기 위해 상당한 시간과 노력이 요구된다. 소리를 매개로 하는 실기 위주의 교과 특성상 텍스트로 구성된 구체적인 채점 기준을 마련하는 것은 좀처럼 쉬운 일이 아니다. 중, 고등학교 음악 교사의 경우, 교사당 담당하는 학생 수가 많기에 짧은 시간 내에 많은 분량의 답안을 평가해야 하는 서·논술형 평가의 채점 과정은 부담이 될 수 있다. 더구나 동일한 평가 문항에 대해 채점자마다 점수가 달라질 수 있으며 채점에서의 신뢰도를 확보하기 어렵다는 점은 서·논술형 평가를 꺼리는 원인이 되기도 한다(Park et al., 2023). 설사 평가 기준이 마련되어 있다고 하더라도 채점자의 주관적 판단이 개입될 수 있기에 평가의 공정성에 문제가 될 수 있다는 것이다(Seong & Shin, 2023). 이러한 까닭에 서·논술형 평가가 오랜 기간 정책적으로 강조되었으나 학교 현장에 깊숙이 뿌리내리지 못했다. 이는 음악 교과에서 서·논술형 평가가 안착하기 위해 문항 개발이나 채점 과정을 지원하는 방안을 모색할 필요가 있음을 시사한다.

이와 관련하여 교육학 분야와 국어, 과학, 지리, 사회, 가정 교과 등을 중심으로 서·논술형 평가에 생성형 AI를 활용하려는 시도들은 주목할 만하다. 이들은 생성형 AI 모델 중 하나인 ChatGPT를 활용하여 서·논술형 평가의 문항을 개발하거나 자동 채점 가능성을 탐색하고자 하였다. ChatGPT는 사전에 학습한 데이터를 기반으로 사용자가 입력한 프롬프트에 따라 새로운 텍스트를 생성해 주며, 최근에는 글쓰기를 위한 아이디어 도출이나 초안 작성, 정교한 추론과 문제 해결에도 강점을 지니는 것으로 보고되고 있다(OpenAI, 2025). 이러한 ChatGPT의 향상된 기능은 사용자 요구에 따라 적절한 응답을 생성함으로써 문항 개발에 대한 교사의 부담을 완화할 수 있을 뿐만 아니라 채점자 간 신뢰도나 채점자 내 신뢰도를 보완할 수 있는 보조적 수단으로 사용될 수 있다(Choi & Park, 2024; Jeong & Baek, 2025; Seong & Shin, 2023). 나아가 ChatGPT를 통한 개인별 맞춤형 피드백은 학생들의 개별화된 학습을 지원함으로써 자기 주도적 학습 능력 신장에도 영향을 줄 수 있을 것으로 기대되고 있다.

이러한 흐름 속에서 음악 교육 분야에서도 생성형 AI의 활용 가능성과 교육적 가치를 모색하려는 연구들이 활발히 이루어져 왔다(Han & Han, 2025; Kim & Hong, 2024; Yoon, 2024). 이들 연구는 생성형 AI가 학생들의 음악적 사고와 표현을 확장하고 교사의 수업 설계를 지원하는 도구로서 활용 범위가 더욱 확대될 거라 주장한다. 하지만 평가와 관련해서는 생성형 AI가 미치는 영향에 대해 이론적으로 논의하거나(Holster, 2024), 설계 방안을 단편적으로 제시하는 데에 그치면서(Ji et al., 2024; Shaw, 2024), 학교 현장에서 서·논술형 평가를 실질적으로 운영하기 위한 구체적인 방안을 실증적으로 조망하는 데에는 한계가 있다.

따라서 본 연구는 생성형 AI를 활용하여 음악 교과 서·논술형 평가 방안을 탐색하는데 목적이 있다. 이를 위해 대표적인 생성형 AI로 알려진 ChatGPT를 통해 중학교 1학년 음악 수업에 적용할 수 있는 서·논술형 평가 문항을 개발하여 실제 평가를 실행한 후 채점하였다. 다만, 이 연구에서는 AI 기반 평가의 효과성을 검증하기보다는 음악 교사들의 서·논술형 평가의 문항 개발과 채점 업무를 덜어주기 위한 보조 도구로서 ChatGPT의 활용 가능성을 살펴보는 데에 초점을 두었다. 이에 따라 다음과 같은 연구 문제를 설정하였다. 첫째, 생성형 AI를 활용하여 음악 교과 서·논술형의 평가 문항을 어떻게 개발할 수 있는가? 둘째, 생성형 AI를 활용하여 음악 교과 서·논술형 평가의 학생 답안을 어떻게 채점할 수 있는가?

II. 이론적 배경

1. 서·논술형 평가의 의미와 중요성

서·논술형 평가라는 용어는 2005년 서울시교육청의 「학업성적관리지침」에 서술형과 논술형을 포괄하여 ‘서·논술형’라고 처음 명시한 이후 학생 평가의 한 방법으로 사용되고 있다. 하지만 서·논술형 평가에 대한 학문적 논의가 이루어지기도 전에 학교 현장에 정책적으로 도입되어 온 까닭에 서술형과 논술형이 혼용되거나 서술형이 논술형의 하위 유형으로 여겨지는 등 개념과 층위에 있어 혼란을 가져왔다(Park et al., 2019). 경우에 따라 제한된 논술형(restricted essay)과 확장된 논술형(extended essay)으로 구분하기도 한다(Gyeonggido Office of Education, 2025).

표준국어대사전에 따르면, 서술(敍述)은 사건이나 생각 따위를 차례대로 말하거나 적는 것을 의미하며, 논술(論述)은 어떤 것에 관하여 의견을 논리적으로 서술하는 것을 뜻한다(Park, 2024). 이러한 정의에 비추어보면, 서술형은 학생이 가지고 있는 지식, 개념, 원리 등을 한두 문장 내외로 작성하는 것으로, 요약, 설명, 비교, 대조, 인과 등을 명확하게 기술하는 데에 초점을 둔다. 이와 달리 논술형은 자신의 주장과 이를 뒷받침할 만한 근거나 이유를 한 문단이나 한편의 글 수준으로 설득력 있게 제시하는 것으로, 제안, 해결, 추론, 비평 등과 같이 다양한 지식과 사고 능력을 복합적으로 활용하는 데에 중점이 있다. 그렇기에 하나의 정답보다 다수의 응답이 가능하며 예외적인 답안을 허용하기도 한다(Seong et al., 2025). 정리해 보면, 서술형 문항은 작성해야 하는 답안의 분량이 비교적 적고 내용의 ‘정확성’에 초점이 있지만, 논술형 문항은 내용 범위나 서술 양식에 있어 응답의 자유도가 허용되며 ‘논리성’을 요구한다는 점에서 다소 차이를 지닌다.

중요한 것은 서·논술형 평가가 학생과 교사 모두에게 긍정적 영향을 줄 수 있다는 점이다(Kim, 2024). 학생들은 자신의 의견을 명확하게 기술하고 논리적으로 사고한 결과를 글로 작성해 보는 과정을 통해 지식을 확장하고 내면화할 수 있다. 교사들은 학생 답안을 통해 학생들의 이해 수준과 학습 상태를 파악함으로써 다음 수업 설계를 위한 구체적인 정보를 수집할 수 있다(Seong & Shin, 2023). 음악 교과에 적용해 보면, 학생들은 음악에 대한 느낌이나 해석을 언어로 구조화하면서 비판적 사고력과 표현력을 함양할 수 있다. 교사들은 학생들의 음악적 성장 정도나 발달 과정을 확인할 수 있으며, 수업의 목표, 내용, 방법을 개선하기 위한 수단으로 활용할 수 있다. 이는 학기 말에 해오던 실기 중심의 일회적인 평가 방식에서 벗어나 학생들의 다양하고 복합적인 능력을 총체적으로 파악하게 해준다. 이러한 맥락에서 음악 교과의 서·논술형 평가는 “음악의 핵심적 지식·이해, 과정·기능, 가치·태도를 포괄하는 다양한 음악 활동을 통하여 감성, 창의성, 자기 주도성을 기르고, 일상생활 속에서 다양한 공동체 안에서 소통할 수 있는 인간 육성”(Ministry of Education, 2022, p. 7)이라는 음악 교과의 목표를 실현하는 데에 중요한 교육적 의미를 지닌다고 볼 수 있다.

2. 교육과정 개정에 따른 서·논술형 평가의 변화

서·논술형 평가에 관한 내용은 6차 교육과정 문서에 처음 언급되었다. 총론에서 교과의 평가는 선택형 중심의 지필평가에 치우치지 않도록 하고, 서술형 주관식 평가와 표현 및 태도의 관찰 평가가 조화롭게 이루어질 필요가 있음을 제시하였다. 음악과 교육과정에서는 서술형 평가라는 용어를 직접적으로 언급하지 않았으나 중·고등학교 감상 영역에서 음악을 듣고 발표하거나 느낌이나 생각을 감상문 형태로 제시함으로써 음악적 정서, 음악 애호심, 음악에 대한 태도 등을 평가할 수 있음을 강조하였다. 이후 7차 교육과정 및 2007 개정 교육과정 시기에는 수행 평가가 본격적으로 도입·적용되면서 학생 평가의 개선을 위해 서술형 평가의 활용이 권장되었다. 2009 개정 교육과정에서는 선택형 중심 평가 방식에서 벗어나 서술형 및 논술형 평가, 수행 평가의 비중을 늘려서 교과별 특성에 적합한 평가를 하도록 하였으며, 서술형과 논술형 문항을 구분하기 시작하였다. 음악과 교육과정에서도 서술형 및 논술형 검사, 발표 및 토론, 연구보고서 등 다양한 평가 방법의 활용을 제안하였다. 2015 개정 교육과정에서는 역량교육의 강조와 함께 ‘과정중심평가’에 대한 관심이 높아지면서 이를 실현하기 위한 평가 방법으로 서·논술형 평가의 중요성이 더욱 부각되었다. 이 시기부터 일부 시도교육청에서는 「학업성적관리지침」을 통해서 서·논술형 평가의 비율과 운영 방침을 제시하고, 교사의 평가 전문성 강화를 위한 연수 프로그램을 운영하거나 평가 예시 문항을 개발하여 학교 현장에 보급하고자 하였다.

이러한 정책적 기조는 2022 개정 교육과정에서도 지속되고 있으며, 총론과 음악과 교육과정 모두 다양한 평가 방법 가운데 하나로 서·논술형 평가를 제시하고 그 비중의 확대를 강조하고 있다. 더불어 2022 개정 교육과정에서는 평가의 방향과 방법을 구체적으로 제시함으로써 교수·학습 과정이나 영역별 특성에 따른 평가 가이드를 상세히 안내하고 있다. 이상과 같이, 6차 교육과정을 시작으로 최근 고시된 2022 개정 교육과정에 이르기까지 서·논술형 평가의 중요성이 점차 강화되고 있다. 따라서 음악 교과에서도 서·논술형 평가 실행을 실질적으로 지원할 수 있는 새로운 도구와 방법에 대한 탐색이 요구된다.

3. 서·논술형 평가에서 생성형 AI의 활용

생성형 AI는 대규모 데이터를 학습하여 텍스트나 이미지와 같은 다양한 형태의 결과물을 생성할 수 있는 인공지능 시스템을 의미한다(Sengar et al., 2024). 교육 분야에서 활발히 논의되고 있는 ChatGPT는 자연어를 사용하여 텍스트를 생성하는 대화용 챗봇으로, 데이터의 패턴과 구조를 학습하여 다음 단어를 예측하는 방식으로 사용자의 질문에 응답한다. ChatGPT는 GPT-1(2018) 이후 끊임없이 발전해 오면서 최근에는 논리적 분석 및 추론, 멀티모달 처리, 코딩 능력 등이 강화된 모델들이 출시되고 있으며 활용 범위가 더욱 확대되고 있다.

이에 따라 평가 분야에서도 ChatGPT를 활용하려는 새로운 시도들이 이루어지고 있다. 이러한 시도들은 학습 목표, 평가 과제, 성취 기준에 근거하여 문항 초안을 생성하거나 채점에 필요한 예시 답안이나 채점 기준표를 설계하는 데에 ChatGPT를 사용하였다(Baek & Park, 2025; Choi & Park, 2024; Ham et al., 2024). 또한 인간 평가자와 ChatGPT의 채점 결과를 분석하는 연구들도 있었다(García-Varela et al., 2025; Im & Lim, 2024; Park et al., 2024; Seong & Shin, 2023; Shin & Yoo, 2024). 이들 연구에서는 평가 영역이나 문항의 유형에 따라 다소 차이가 있으나, 채점 기준이 비교적 명확하게 제시된 항목에서 ChatGPT 채점 결과가 교사 채점과 유사한 수준의 정확성을 보이는 것으로 확인되었다. 이는 교사의 채점 업무를 보조하는 도구로서 ChatGPT의 활용 가능성을 시사한다.

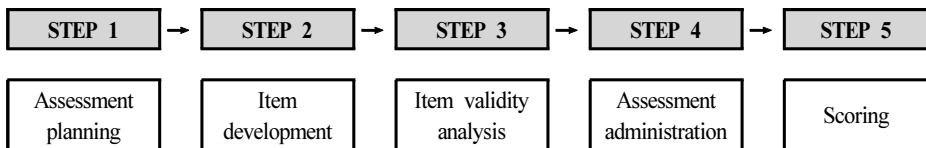
한편, ChatGPT를 효과적으로 활용하기 위해 프롬프트 엔지니어링을 적용한 연구들도 수행되고 있다. 이들 연구에 따르면, 최소한의 정보만 제공하는 단일 명령어 방식의 ‘Zero-Shot’ 프롬프트보다 구체적인 맥락과 여러 개의 예시를 포함하는 ‘Few-Shot’ 프롬프트가 안정적인 응답을 도출하는 데 효과적인 것으로 나타났다(Brown et al., 2020; Kwon, 2024; Moore et al., 2022). 최근에는 답변을 생성하는 과정에서 단계별 추론을 유도하는 ‘CoT’ 프롬프트가 새롭게 시도되고 있다(Jeong & Baek, 2025; Park & Choi, 2025; Wei et al., 2023). 이러한 방식은 별도의 미세 조정(fine-tuning)이나 모델에 대한 추가 학습 없이도 채점이

가능하다는 점에서(Kwon, 2024; Shin & Yoo, 2024), 생성형 AI 활용에 익숙하지 않은 교사들도 비교적 쉽게 적용할 수 있는 것으로 평가된다.

종합해 보면, 생성형 AI는 교사의 평가 업무를 지원하고 교수·학습 및 평가의 질을 제고하는 데 도움을 제공함으로써 서·논술형 평가의 실행 과정에서 제기되어 온 문제들을 보완할 수 있을 것으로 보인다. 다만, 잘못된 정보로 그럴듯하게 답변하는 환각 현상(hallucination), 편향적 응답, 채점의 일관성 문제 등은 여전히 중요한 과제로 남아있다(Swiecki et al., 2022). 따라서 생성형 AI가 도출한 문항의 타당성과 채점 결과의 신뢰성을 확보하기 위한 체계적인 검증과 보완 과정이 필요하다.

III. 연구 방법 및 절차

본 연구는 생성형 AI를 활용한 음악 교과 서·논술형 평가 방안을 탐색하기 위해 ChatGPT를 통해 중학교 1학년 음악 수업에 적용할 수 있는 서·논술형 평가 문항을 개발하여 실제 평가를 실행한 후 채점하였다. 연구 기간은 2025년 3월~10월까지이며 3월~5월에는 문헌 검토와 연구 설계, 6월~10월에는 자료 수집 및 분석이 이루어졌다. 실질적인 연구 절차는 ‘평가 계획’-‘문항 개발’-‘문항 타당성 검토’-‘평가 실행’-‘채점’의 단계로 진행되었으며, ‘문항 개발’과 ‘채점’ 단계에서 ChatGPT를 활용하였다. 단계별 구체적인 내용은 다음과 같다.



[Figure 1] Research design and process

1. 평가 계획

평가 계획 단계에서는 영역, 성취 기준, 학습 목표를 선정하고, 이를 토대로 평가 과제를 도출하였다. 영역은 2022 개정 음악과 교육과정에서 대영역으로 새롭게 분리된 ‘창작’에 기반하였다. 이는 음악사나 감상 중심의 제한된 영역에서 벗어나 다양한 맥락에서 서·논술형 평가의 적용 가능성을 탐색하기 위함이다. 성취 기준은 ‘음악적 의도나 아이디어를 여러 매체나 방법에 적용하여 자기 주도적으로 창작한다[9음03-01]’, ‘음악의 요소

와 특징을 활용하여 간단한 형식의 음악을 만든다[9음03-03]', '생활 속의 영역과 연계하여 음악을 만들고 활용하며 책임감을 갖는다[9음03-04]'를 선정하였다. 학습 목표는 '악기의 구조 및 소리 나는 원리를 파악하고 자신만의 악기를 만들 수 있다'와 '자신이 만든 악기로 학교 종소리에 어울리는 음악을 창작할 수 있다'로 구성하였다.

또한 수업 시간에 학습한 내용을 바탕으로 주체적으로 문제를 해결해 가는 6차시의 프로젝트 학습을 통해 과정에 대한 평가가 이루어지도록 하였다. 1차시에는 악기의 구조와 소리 나는 원리에 대해 탐색하고 선택한 악기에 대해 조사·발표하였다. 2~3차시에는 메이키메이키 보드와 엔트리 코딩 프로그램을 활용하여 악기를 제작하였다. 4차시에는 기존 학교 종소리를 함께 감상하면서 리듬, 가락, 셈여림, 빠르기 등의 음악 요소와 특징을 분석하였다. 5~6차시에는 제작한 악기를 활용하여 1분 이내의 학교 종소리를 만들었다. 이상의 내용을 바탕으로, 평가 과제는 구체적인 문제 상황과 학습 경험을 포함하여 '자신이 만든 악기로 학교 종소리 창작하기'로 설정하였다. 이를 통해 실생활 맥락 속에서 학습한 개념이나 원리를 적용하여 간단한 형식의 음악을 만들 수 있는지를 평가하고자 하였다.

2. 문항 개발

문항 개발 단계에서는 ChatGPT 4.5 모델을 바탕으로 ChatGPT 3.5와 ChatGPT o3-pro 모델을 병행하여 서·논술형 평가 문항과 문항별 채점 기준을 생성하였다. 이때 채점 기준은 평가 요소별 수행 수준과 배점이 포함된 구체적인 루브릭의 형태로 제작하였다. 즉 문항별로 세 가지 평가 요소로 구분하고 상, 중, 하의 수행 수준에 따라 5점, 3점, 1점을 부여하였다. 이 단계에서는 역할 부여하기, 요구와 의도 정확하게 전달하기, 관련 자료 첨부하기와 같은 프롬프트 전략과 함께 평가 계획, 서술형과 논술형 문항의 의미와 특징, 차시별 수업 내용 및 활동지를 PDF 파일로 정리하여 활용하였다.

3. 문항 타당성 검토

문항 타당성 검토 단계에서는 ChatGPT가 생성한 평가 문항 및 채점 기준에 대한 전문가 검토를 진행하였다. 전문가는 대학교수 1명과 교사 4명, 총 5명으로 구성하였다. 이때 교사 전문가는 1) 10년 이상의 교직 경력, 2) 교육과정, 수업, 평가 관련 연수 강사로서 활동 이력, 3) 5년 이상 서·논술형 평가 실행 경험을 지닌 중, 고등학교 음악 교사로 선정하였다.

<Table 1> Teacher participant's information

Participant	School level	Teaching experience	Major
Teacher A	Middle school	25 years	Western music (Piano)
Teacher B	Middle school	12 years	Western music (Vocal)
Teacher C	Middle school	12 years	Western music (Piano)
Teacher D	High school	10 years	Western music (Piano)

전문가들은 평가 문항 및 채점 기준이 타당한지를 1점(전혀 타당하지 않다)에서 4점(매우 타당하다)까지 리커트 척도로 평정하였으며, 개선 사항이나 기타 의견을 자유롭게 기술하여 회신하였다. 타당성은 평균(M)이 3.00 이상이고 표준편차(SD)가 0.80 이하일 때에 적절하다고 판단하였다. 다만, 타당성 검토에 참여한 전문가 수가 제한적이라는 점을 고려하여, 평균이나 표준편차와 같은 양적 수치뿐 아니라 이들이 제시한 의견을 중심으로 검토하였다. 또한 일부 교사들과 약 1시간 내외의 심층 면담을 진행하여 평가 문항 및 채점 기준에 대한 사항 외에도 ChatGPT의 활용 경험에 관해 다양한 이야기를 나누었다. 이를 통해 평가 문항의 오류나 채점 기준의 적합성을 점검하고 실제 평가 상황에서 발생할 수 있는 문제를 반영하여 최종 평가 문항 및 채점 기준을 개발하였다.

4. 평가 실행

평가는 A 중학교 1학년, 6개 학급, 학생 150명(재적 161명 중 특수학급 및 위탁학생 포함 결석생 11명 제외)을 대상으로 2025년 6월 9일~13일까지 5일 동안 실행되었다. 실제 시험은 정규 음악 수업 시간에 서·논술형 문항이 탑재된 웹 기반 환경에서 45분 동안 이루어졌다. 이 과정에서 답안 작성 및 수집을 위해 경기도교육청이 운영하는 하이러닝 교수·학습 플랫폼을 사용하였다. 해당 플랫폼은 교사가 평가 문항을 제시하고 학생들이 답을 입력하도록 설계되어 있으며 자동 채점과 피드백 기능을 제공한다. 하지만 연구를 수행할 당시 음악 교과에는 이 기능이 지원되지 않았다. 이에 따라 시스템상 국어 교과로 설정한 후에 문항과 채점 기준을 탑재하고, ‘학습 콘텐츠’ 메뉴에 포함된 ‘AI 서·논술형 출제’ 기능을 통해 학생 답안을 수집 및 데이터화하는 목적으로 사용하였다.

한편, 본 연구를 수행할 당시 A 중학교의 1학년 1학기는 자유학기제로 운영되고 있어 교과 성취도 점수를 산출하는 대신 학생의 성장 및 발달에 대한 평가 결과를 학교생활기록부에 문장 형식으로 기록하게 되어 있었다. 이에 따라 ChatGPT를 통해 산출된 점수는 실제 성적에 반영하지 않고 학습 목표의 달성 정도를 점검하기 위한 연구 자료로만 활용하였다.

5. 채점

채점 단계에서는 서·논술형 평가에 참여한 6개 학급 중 임의로 한 학급을 선정하여 25명 학생의 답안을 분석 대상으로 삼았다. 수집된 답안은 서술형과 논술형으로 구분하여 문항별로 25개, 총 50개를 엑셀 파일로 변환하고 이름, 학급, 번호와 같은 개인 식별 정보는 삭제하였다. 먼저, ChatGPT 채점은 프롬프트 엔지니어링과 맞춤형 GPT 방식을 활용하였다. 이때에는 ChatGPT 5 모델을 사용하였는데, 이는 문항 개발에 사용했던 기존 모델의 사용이 어려워지면서 다수의 학생 답안을 안정적으로 처리할 수 있는 모델이 필요했기 때문이다. 이어서, 교사 채점은 A 중학교에 근무하는 음악 교사 2명이 채점 결과(평균, 표준편차, 최솟값, 최댓값)를 공유하면서 각자 점수를 부여하였다. 끝으로, ChatGPT와 교사 채점 결과를 비교하기 위해 피어슨 상관계수(Pearson's r)를 산출하였으며, .60 이상 일 때 강한 상관관계로 해석하여 채점자 간 신뢰도가 높은 것으로 판단하였다(Seong et al., 2025).

IV. 연구 결과

1. 생성형 AI를 활용한 문항 개발

서·논술형 평가 문항을 개발하기 위해 ‘중학교 1학년 음악 교사’라는 역할을 부여하고 평가 과제(자신이 만든 악기로 학교 종소리 창작하기)와 문항 유형(서술형과 논술형 문항 각 1개씩)을 포함하여 프롬프트를 구성하였다. 또한 배경지식과 맥락 정보를 제공하기 위해 평가 계획, 서술형과 논술형 문항의 의미와 특징, 차시별 수업 내용 및 활동지를 PDF 파일로 첨부하였다. 이때 평가 계획에는 2022 개정 음악과 교육과정에 기반하여 영역, 성취 기준, 학습 목표 등을 포함함으로써 교육과정과의 관계 속에서 문항을 개발하고자 하였다. 이상의 내용을 바탕으로 문항 개발에 사용한 프롬프트는 다음과 같다.

<Table 2> Prompt for item development

Prompt
Assume the role of a first-year middle school music teacher. For the assessment task “Creating a School Bell Using a Self-Made Instrument,” develop one restricted essay item (15 points) and one extended essay item (15 points). Refer to the attached materials, including the assessment plan; definitions and characteristics of restricted and extended essay items, and worksheets for each class session.

그 결과, 동일한 프롬프트를 제시했음에도 질문할 때마다 ChatGPT가 생성한 문항의 내용이나 범위에 차이가 있었다. 예컨대 서술형 문항의 경우, 세 차례의 응답 모두 ‘악기의 구조와 소리 나는 원리’를 포함했으나 ‘악기 제작 과정에서 고려한 음악 요소(응답 2)’나 ‘낼 수 있는 소리나 음높이(응답 3)’와 같은 내용이 서로 다르게 나타났다. ChatGPT는 사전에 학습된 데이터와 프롬프트를 바탕으로 확률적으로 가장 많이 나타나는 언어를 추론하기 때문에 어떠한 단어와 문장 구조를 선택하느냐에 따라 다양한 답변을 생성하기 때문이다. 논술형 문항의 경우, 평가 계획에 포함된 세 가지 성취 기준 중 어디에 중점을 두는지에 따라 범위가 달라졌다. 즉 응답 1은 [9음03-01]의 ‘음악적 의도나 아이디어’, 응답 2는 [9음03-03]의 ‘음악의 요소와 특징’, 응답 3은 [9음03-04]의 ‘생활 속 활용’에 초점을 두고 문항을 생성했다. 이는 ChatGPT가 아무리 많은 데이터를 동시에 처리할 수 있을지라도 하나의 프롬프트에 다수의 자료를 한꺼번에 포함하기보다는 짧고 단순한 지시문을 통해 의도를 명확하게 전달할 필요가 있음을 보여준다.

<Table 3> Restricted and extended essay items generated by ChatGPT

Restricted essay	Response 1	Write the name of the instrument you created and describe in detail its structure and the principle by which it produces sound.
	Response 2	Explain the structure and the principle of sound-producing principle of the instrument you created, and describe specifically which musical elements you considered during the process of making the instrument.
	Response 3	State the name, structure, and sound-producing principle of your instrument, and describe the type of sounds or pitches this instrument can produce.
Extended essay	Response 1	In the process of creating a school bell with your instrument, identify the most important musical intention or idea you considered, and present your reasons logically.
	Response 2	Discuss how you applied musical elements and characteristics when composing a school bell with your instrument, and explain your reasoning logically.
	Response 3	Explain why you think the instrument you created is more effective than the existing school bell, and provide specific examples of how this instrument could be utilized in school setting.

한편, ChatGPT 모델의 유형에 따라 응답에 차이를 보이기도 했다. 낮은 버전의 무료 모델인 ChatGPT 3.5와 달리 최신 버전의 유료 모델인 ChatGPT 4.5나 ChatGPT o3-pro의 경우 ‘음높이’, ‘리듬’, ‘음악 요소나 음악적 특징’과 같은 특정 용어에 대해 부연 설명을 덧붙이거나 구체적으로 서술하는 것으로 나타났다. 특히 ChatGPT o3-pro는 줄의 수를 4-5줄 이내로 제한하거나 문항을 구조화하여 제시하였으며, 별표(**) 사이에 텍스트를 넣어 일부 단어를 강조하기도 하였다.

<Table 4> Restricted and extended essay items according to ChatGPT models

Type	ChatGPT 3.5	ChatGPT 4.5	ChatGPT o3-pro
Restricted essay	Briefly describe the structure of the instrument you created, the materials used, and the principle by which it produces sound.	State the name, structure, and sound-producing principle of the instrument you created in detail, and write down the possible musical scales or pitches (e.g., Do, Re, Mi) that can be played with this instrument.	Write the name and the main materials of the instrument you created, and explain the principle (aerophone, chordophone, membranophone, idiophone, etc.) by which it produces sound. Then, in 4-5 sentences, describe how you selected the **pitches (high/low)** and **rhythms (length/intensity)** to represent the school bell.
Extended essay	Compare your self-created school bell with the existing school bell, explain the characteristics and creative intention of your composition, and logically describe the positive influences this bell could have on school life.	When composing a school bell with your self-created instrument, explain which musical elements or characteristics (e.g., rhythm, pitch, dynamics, timbre) you considered. Then, logically describe your thoughts on the atmosphere or message you intended to convey through the composition.	Reflect on the process of creating and performing a school bell with your instrument. 1. Describe in detail the most impressive experience or the most difficult challenge you encountered. 2. Explain logically how this activity contributed to your musical creativity, sense of responsibility, and ability to utilize digital resources.

흥미로운 점은 별다른 요청을 하지 않았음에도 불구하고 세 모델 모두 채점 기준을 함께 생성했다는 것이었다. 그러나 처음 ChatGPT가 생성한 채점 기준은 평가 요소와 요소별 배점을 단편적으로 제시하는 수준에 그쳤다. 이에 따라 구체적인 채점 기준을 도출하기 위해 ‘#’으로 문장을 구분하고 목적(command), 제한 사항(constraints), 출력 조건(output)을 단계별로 제시하였다. 즉 목적에는 문항별 채점 기준을 상세하게 제시하고, 제한 사항에는 교육과정에 제시된 내용 요소를 채점 기준에 반영하며, 출력 조건에는 ‘~할 수 있다’의 형식을 사용하여 표를 만들어달라고 요청하였다. 이 과정에서 ChatGPT는 종종 프롬프트를 완전히 반영하지 못한 답변을 하기도 했다. 예를 들어, 채점 기준의 언어 형식(~할 수 있다)을 따르지 않거나 내용 요소에 제시된 지식·이해, 과정·기능, 가치·태도를 채점 기준에 제대로 반영하지 못했다. 이에 따라 내용 요소에 대한 설명이 포함된 관련 자료를 추가로 제공하고, ChatGPT가 잘못된 답을 할 때마다 이를 지적하는 과정을 반복하였다. 또한 ‘CoT’ 프롬프트 방식을 적용하여 어떠한 과정을 통해 답변을 도출했는지 근거를 제시하도록 하였다. 그리고 나서 한글 맞춤법에 어긋나거나 어색한 문장들은 문체에 맞도록 수정한 후 평가 문항과 채점 기준(초안)을 도출하였다.

이렇게 도출된 문항의 타당성 검토 결과, 3.00~3.80의 평균값을 보였는데, 8개 항목에서 ‘매우 타당하다’ 혹은 ‘약간 타당하다’에 응답한 비율이 100%에 이를 정도로 높은 수준을 보였다. 표준편차 역시 최대 0.55로 나타나 전문가들 간 의견의 차이가 크지 않음을 확인할 수 있었다. 다만, 서술형 문항의 평가 요소 3에 해당하는 ‘음의 구성 기술’ 항목에서 가장 낮은 점수($M = 3.00$)를 보이면서 상대적으로 타당도가 낮게 나타났다.

<Table 5> Validity analysis of the assessment tool

Type	Component		Response rate (%)				M(SD)
			Not at all valid	Somewhat not valid	Somewhat valid	Very valid	
Restricted essay	Item		0	0	80	20	3.20(0.45)
	Scoring criteria	Instrument name and structure	0	0	40	60	3.60(0.55)
		Explanation of sound-producing principle	0	0	60	40	3.40(0.55)
		Explanation of pitch composition	0	0	100	0	3.00(0.00)
Extended essay	Item		0	0	20	80	3.80(0.45)
	Scoring criteria	Explanation of application of musical elements	0	0	80	20	3.20(0.45)
		Statement of creative intention and message	0	0	60	40	3.40(0.55)
		Logical organization and completeness	0	0	20	80	3.80(0.45)

한편, 타당성 검토에 참여했던 교사들은 ChatGPT의 뛰어난 성능에 대해 놀라움을 표현했다. 예를 들어 C 교사는 기존에 생각하지 못했던 문항에 대한 새로운 아이디어를 제공하고 짧은 시간 안에 출제 의도를 인지하여 문항을 생성해 준다는 점에서 긍정적인 반응을 보였다. 또한 차시별 활동지와 관련 자료를 제공하면 수업 내용과 관련지어 문항과 채점 기준을 만들어주는 게 놀라웠다고 C 교사는 언급했다. 평소 루브릭을 만들 때 어떠한 단어를 사용할지가 고민이었는데 교육과정에서 제시된 성취 기준에 맞게 평가 요소별로 채점 기준을 개발해 주는 점이 신기했다는 것이다.

반면, 몇몇 교사들은 ChatGPT가 음악 교과 고유의 특성을 제대로 반영하지 못해 학문적 오류가 있는 내용을 포함할 여지가 있다고 문제를 제기하기도 하였다. 특히 서술형 문항 중 평가 요소 3에 해당하는 ‘음의 구성 기술’ 항목에서 수정이 필요하다는 의견이 많았다.

학생이 제작한 악기가 무얼 타악기이면 낼 수 있는 소리를 음높이로 표현하기 힘들지 않나요? 음의 구성(예: 도, 레, 미 등) 또는 리듬(예: 1/2박, 1박, 2박 등)이라고 수정하는 건 어떨까요? 아니면 음과 소리를 같이 적는 것도 괜찮을 거 같아요(A교사).

악기가 소리낼 수 있는 음을 차례대로 기술하는 것이 학생들에게는 어렵게 느껴질 수 있어요. 자신이 만든 악기가 음정 표현이 명확하지 않은 경우, 객관적 평가가 어려울 수 있을 것으로 보입니다. 음의 구성에 어떤 요소가 있는지 안내 문구가 필요합니다(D교사).

음악 교과의 개념에 맞게 음 혹은 소리 기술이라고 표현하는 것이 낫지 않을까요? 음높이
이나 리듬, 음색 같은 요소들을 예로 제시하면 좋을 듯 해요(B교사).

그 외에도 일부 평가 요소의 내용이 부정확하다는 점, 요소별 최저점의 합계가 만점의 40% 미만이 되도록 배점을 조정해야 한다는 점 등의 의견이 있었다. 이를 바탕으로 평가 문항과 채점 기준을 최종 수정 및 보완하였다. 이에 따라 서술형 문항은 자신이 만든 악기의 이름과 구조, 소리 나는 원리를 적고, 이 악기로 낼 수 있는 음 혹은 소리(예: 음높이, 음역, 리듬, 음색 등)에 대해 구체적으로 서술하도록 하였다. 채점 기준은 ‘악기명 및 구조 설명’, ‘소리 나는 원리 서술’, ‘음 혹은 소리 기술’의 세 가지 평가 요소로 구분하고 평가 요소별 수행 수준 및 배점을 부여하였다.

<Table 6> Restricted essay item and rubric

Item	Write the name of the instrument you made. Describe its structure and how it produces sound. Also, explain what kinds of sounds or notes(for example, pitch, range, rhythms, or timbre) this instrument can make.	
Criteria	Level of achievement	Points
Identifies the instrument and describes its structure	Provides a clear and detailed identification of the instrument and its structure	5
	Provides a general identification of the instrument and its structure	3
	Provides a limited or vague identification of the instrument and its structure	1
Explains the sound-producing principle	Clearly explains the sound-producing principle in close relation to the instrument's structure	5
	Explains the sound-producing principle in relation to the instrument's structure	3
	Explain the sound-producing principle with little or no relation to the instrument's structure	1
Description of the pitches or sounds	Clearly describes the pitches or sounds produced by the instrument	5
	Describes the pitches or sounds produced by the instrument	3
	Describes the pitches or sounds produced by the instrument in a limited way	1

논술형 문항은 자신이 만든 악기로 창작한 학교 종소리의 음악적 특징을 <조건>에 따라 분석하고, 이를 통해 어떤 의도나 메시지를 표현했는지 자기의 생각을 논리적으로 서술하도록 하였다. <조건>은 음악 요소(리듬, 음높이, 셈여림, 음색) 중 두 가지 측면에서 음악적 특징을 분석하여 500자 이내로 작성해야 한다고 제시하였다. 채점 기준은 ‘음악적 특징 분석’, ‘창작 의도 및 메시지 표현’, ‘논리적 구성 및 완성도’의 세 가지 평가 요소로 구분하고 평가 요소별 수행 수준 및 배점을 부여하였다.

<Table 7> Extended essay item and rubric

Item	Analyze the musical characteristics of a school bell composition created with your self-made instrument according to the conditions below. Then, logically describe your thoughts on the intention or message you wanted to express through the composition. Conditions: Analyze the musical characteristics using two of the following musical elements: rhythm, pitch, dynamics, and timbre. (Up to 500 characters)	
Criteria	Level of achievement	Points
Analysis of musical characteristics	Provides a detailed and accurate analysis of two musical elements	5
	Analyzed musical elements but provides limited or vague explanation	3
	Provides little or no analysis of musical elements, or presents only superficial descriptions	1
Expression of creative intention or message	Clearly and convincingly expresses the creative intention or message in close connection with the musical characteristics	5
	Expresses the creative intention or message but lacks clarity or depth	3
	Provides a vague or insufficient expression of the intention or message	1
Logical organization and completeness	Presents ideas logically with a well-organized structure and minimal grammatical errors	5
	Presents ideas with a generally organized structure, though some grammatical errors are found	3
	Presents ideas with am poorly structure, frequent grammatical errors, and limited logical development	1

2. 생성형 AI를 활용한 채점 결과

서·논술형 평가의 학생 답안을 채점하기 위해 ‘중학교 1학년 음악 교사’라는 역할을 부여하고 평가 문항, 채점 기준, 예시 답안을 포함하여 ‘Few-Shot’ 방식의 프롬프트를 구성하였다. 예시 답안은 교사의 가채점 결과 점수대별 답안 3개를 임의로 선정하였다. 또한 평가 요소별로 채점 후에 채점 결과 및 근거를 엑셀 파일로 제공해달라고 요청하였다. 이는 중간 단계의 사고 과정과 결과를 함께 출력하게 하는 ‘CoT’ 프롬프트에 해당하

는 것이라 볼 수 있다. 이상의 내용을 바탕으로 채점에 사용한 프롬프트는 다음과 같다.

<Table 8> Prompt for scoring

Prompt
Assume the role of a first-year middle school music teacher. Evaluate the students' responses according to the provided evaluation questions, scoring criteria, and sample answers. Assign scores for each evaluation element, calculate the total score, and record the results in an Excel file.

위의 프롬프트를 바탕으로 문항별로 25개의 답안에 대해 ChatGPT 채점을 수행하였다. 이 과정에서 한 번에 많은 양의 프롬프트와 자료를 입력하면 실행이 지연되거나 중단되는 오류가 생기기도 했다. 이에 따라 문항별로 학생 답안을 절반씩 나누어 요청하였다. 더불어 동일한 프롬프트이더라도 질문할 때마다 ChatGPT 응답에 차이가 있다는 점을 고려하여, 3회에 걸쳐 채점하였다. 프롬프트 엔지니어링 방식을 통해 ChatGPT 채점을 수행한 결과는 다음과 같다.

<Table 9> Scoring results of ChatGPT by repeated measures

Type	Criteria	ChatGPT(1st)				ChatGPT(2nd)				ChatGPT(3rd)			
		<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max
Restricted essay	1	4.44	1.23	1	5	4.84	0.54	3	5	4.52	1.33	1	5
	2	3.24	2.03	1	5	2.76	1.82	1	5	3.32	0.95	1	5
	3	2.84	0.80	1	5	2.28	1.69	1	5	3.56	0.92	3	5
	Total score	10.52	2.66	3	15	9.88	2.60	5	15	11.40	1.83	7	15
Extended essay	1	3.80	1.39	1	5	3.52	1.53	1	5	2.92	2.04	1	5
	2	3.24	1.03	1	5	3.52	1.81	1	5	3.56	1.96	1	5
	3	3.64	1.09	1	5	4.26	1.48	1	5	4.68	0.75	3	5
	Total score	10.68	2.63	3	15	11.30	4.10	3	15	11.16	3.26	5	15

그러나 이러한 방식은 채점을 요청할 때마다 관련 자료를 업로드하고 지시 사항을 새롭게 요청해야 하는 번거로움이 따랐다. 한 세션 내에서 사용할 수 있는 데이터가 제한되어 있었기 때문에 오류가 날 때마다 새로운 세션에서 시작하거나 프롬프트를 다시 입력해야 할 때도 있었다. 더 큰 문제는 점수대별 예시 답안을 제공했음에도 불구하고 외부의 데이터에 의한 추론이 함께 이루어지면서 일관된 채점 결과를 얻기가 어려웠다는 점이었다. 그 결과, 회차별 총점 평균은 서술형 문항의 경우 10.52점, 9.88점, 11.40점, 논술형 문항의 경우 10.68점, 11.30점, 11.16점으로, 문항별로 1.52점(서술형), 0.62점(논술형)의 차이를 보였다. 또한 동일한 유형의 문항 내에서도 평가 요소에 따라 적게는 0.32점, 많게

는 1.28점 차이가 나타났다.

이러한 문제를 보완하기 위해 ChatGPT의 GPTs 기능을 활용하여 채점의 효율성과 일관성을 높이고자 하였다. GPTs는 사용자의 요구에 따른 맞춤형 GPT로서, 매번 프롬프트를 입력하지 않아도 반복적으로 작업을 수행하며 제공된 자료에 기반하여 일관된 응답을 도출하는 것으로 알려져 있다. 이에 따라 음악 교과 서·논술형 평가 채점에 활용할 수 있는 맞춤형 GPT를 개발하기 위해 다음과 같이 설정하였다.

GPT의 이름은 ‘평가 채점 도우미’로 입력하고, ‘음악 교과의 서술형과 논술형 평가의 학생 답안을 채점해 주는 GPT입니다’라는 설명을 덧붙였다. 지침(instructions)에는 ‘Knowledge에 업로드된 채점 기준과 예시 답안을 참고하여 채점할 것’, ‘평가 요소별로 채점하고 채점 근거와 함께 총점을 산출할 것’, ‘채점 결과는 엑셀 파일로 제공할 것’이라고 입력하였다. 지식창고(knowledge)에는 GPT가 작업을 수행하는 데에 참고하도록 평가 문항, 채점 기준, 예시 답안, 수업 관련 활동지, 2022 개정 음악과 교육과정 문서 등을 등록하였다. 그리고 문항 타당도가 상대적으로 낮게 나타났던 서술형 문항 중 평가 요소 3의 채점 기준을 보완하기 위해 음악적 개념과 요소에 관한 내용이 담긴 음악 이론 자료를 추가하였다. 대화 스타터(conversation starters)에는 GPT 사용 시 처음 시작하는 문장으로 ‘채점해 줘’라고 입력하였다. 기능(capabilities)에는 제공된 자료와 규칙 내에서만 작업하도록 웹페이지 검색 기능을 해제하고, 파이썬(Python)으로 정보를 분석하여 답변을 제공하는 코드 인터프리터 기능을 활성화하였다.

<Table 10> Configure of customized GPT

	Content
Name	Evaluation grading assistant
Description	A GPT designed to evaluate students' responses to descriptive and extended essay-type questions in music education.
Instructions	- Apply the grading criteria strictly based on the provided sample answers. - Evaluate according to each assessment element and present supporting evidence. - Provide the grading results in an Excel file. - Assign the Excel file name in English.
Knowledge	Questions, grading criteria, sample answers, class activities, and related textbook
Conversation starters	"Grade it"
Capabilities	☐ Web browsing ☒ Code interpreter

맞춤형 GPT를 활용하여 채점을 수행한 결과는 다음과 같다<Table 11>. 이러한 결과는 <Table 9>와 같이 프롬프트 엔지니어링만으로 채점했던 방식과 달리 반복적인 측정에도 채점 결과가 달라지지 않았다.

<Table 11> Scoring results of customized GPT

Type	Criteria	Customized GPT			
		<i>M</i>	<i>SD</i>	Min	Max
Restricted essay	1	3.96	1.10	3	5
	2	4.28	1.14	1	5
	3	3.32	0.95	1	5
	Total score	11.56	1.69	5	15
Extended essay	1	3.67	0.96	3	5
	2	4.42	0.93	3	5
	3	3.25	1.23	1	5
	Total score	11.33	1.83	7	15

한편, <Table 12>에서 제시된 바와 같이 동일한 답안에 대해 교사 채점을 수행한 결과, 서술형 문항에서 교사 1의 평균은 7.08점으로 교사 2의 평균 6.44점보다 0.64점이 높았다. 반면 논술형 문항에서 교사 1의 평균은 8.28점으로 교사 2의 평균 8.84점보다 0.56점이 낮았다. 평가 요소별 점수를 살펴보면, 서술형 문항의 평가 요소 3과 논술형 평가의 평가 요소 3에서 다른 평가 요소에 비해 교사 간 점수 차이가 상대적으로 크게 나타났다. 그러나 최솟값과 최댓값을 비교한 결과, 서술형 문항의 평가 요소 3에 해당하는 항목의 최댓값 3점을 제외하고는 교사 1과 교사 2의 채점 결과는 동일하였다. 또한 교사 간 채점 결과에 대한 *t* 검증 결과, 서술형 및 논술형 평가 모두 교사 1과 교사 2의 평균 점수 차이가 통계적으로 유의하지 않았다. 이는 교사 간 평균 점수에 통계적으로 유의한 차이가 나타나지 않았음을 의미한다.

<Table 12> Comparison of scoring results across teachers

Type	Criteria	Teacher 1				Teacher 2				<i>t</i>	<i>p</i>
		<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max		
Restricted essay	1	3.08	1.32	1	5	2.84	1.69	1	5	0.56	.579
	2	2.44	1.33	1	5	2.36	1.47	1	5	0.20	.841
	3	1.56	1.06	1	5	1.24	0.65	1	3	1.29	.206
	Total score	7.08	2.80	3	15	6.44	3.13	3	13	0.76	.449
Extended essay	1	3.32	1.35	1	5	3.40	1.50	1	5	-0.20	.843
	2	2.76	1.42	1	5	2.92	1.55	1	5	-0.38	.705
	3	2.20	1.60	1	5	2.52	1.17	1	5	-0.81	.424
	Total score	8.28	3.96	3	15	8.84	3.71	3	15	-0.52	.608

p* < .05, *p* < .01, ****p* < .001

더불어, 교사 채점 결과를 추가로 제공한 후 맞춤형 GPT를 활용하여 채점을 다시 수행하였다. 분석 결과, 서술형 문항의 평균 점수는 6.76점, 논술형 문항의 평균 점수는 8.56점으로, 교사 채점 결과를 제공하기 전<Table 11>에 비해 각각 4.80점, 2.77점 낮게 나타났다. 최솟값과 최댓값의 경우, 교사 채점 결과를 제공하기 전과 비교했을 때 서술형 문항의 최솟값은 5점에서 3점으로, 최댓값은 15점에서 13점으로 낮아졌다. 또한 논술형 문항의 최댓값은 변화가 없었으나 최솟값은 7점에서 3점으로 낮아졌다. 이는 교사 채점 결과를 추가한 이후 맞춤형 GPT가 이전보다 엄격한 기준으로 점수를 부여하는 경향이 있음을 보여준다.

<Table 13> Scoring results of customized GPT+teachers

Type	Criteria	Customized GPT+Teacher			
		M	SD	Min	Max
Restricted essay	1	2.96	1.40	1	5
	2	2.40	1.19	1	5
	3	1.40	0.71	1	3
	Total score	6.76	2.80	3	13
Extended essay	1	3.36	1.35	1	5
	2	2.84	1.41	1	5
	3	2.36	1.19	1	5
	Total score	8.56	3.71	3	15

마지막으로, ChatGPT의 채점 방식(맞춤형 GPT, 맞춤형 GPT+교사 채점)에 따른 평가 결과를 교사 채점과 비교하였다. 이때 교사 채점은 교사 2인의 평균 점수를 산출하여 사용하였다.

<Table 14> Comparative analysis of ChatGPT and teachers scoring results

Type	Restricted essay	Extended essay
Customized GPT	.070	.580
Customized GPT+Teacher	.961	.800

총점 간 상관계수를 살펴보면, 서술형 문항의 경우 맞춤형 GPT와 교사 채점 간의 상관계수는 .070으로 통계적으로 유의미한 상관관계가 나타나지 않았다. 반면, 교사 채점 결과를 포함한 맞춤형 GPT와 교사 채점 간의 상관계수는 .961로서 매우 높은 수준의 상관관계를 보였다. 논술형 문항의 경우 맞춤형 ChatGPT와 교사 채점 간의 상관계수는 .580으로

로 중간 이상의 상관관계를 나타냈으며, 교사 채점 결과를 포함한 맞춤형 GPT와 교사 채점 간의 상관관계수는 .800으로 강한 상관관계가 확인되었다.

V. 논의 및 결론

본 연구는 음악 교과 서·논술형 평가에 있어 음악 교사들이 겪고 있는 어려움을 해결하기 위한 하나의 대안으로 생성형 AI에 주목하여 문항 개발과 채점 단계에서 ChatGPT의 활용 가능성을 탐색하였다. 특히 이 연구에서는 ChatGPT를 다양한 방식으로 적용하고 그 결과를 비교·분석함으로써 보다 효과적인 활용 방안을 도출하는 데에 초점을 두었다. 주요 연구 결과를 정리하고 이에 대한 논의를 제시하면 다음과 같다.

첫째, 문항 개발 단계에서는 동일한 프롬프트를 제시할 때마다 ChatGPT가 생성한 문항의 내용이나 범위가 달라지는 것으로 나타났다. 또한 상위 버전 모델일수록 음악 용어에 대해 부연 설명을 덧붙이거나 문항을 구조화하여 제시하는 경향을 보였다. 가령, 최신 유료 모델의 경우 답안 길이를 제한하거나 핵심 개념을 강조하면서 학습자 수준을 고려하여 문항을 생성한다는 점을 확인할 수 있었다. 이러한 결과는 서·논술형 문항 개발에 있어 ChatGPT의 효과를 검토한 여러 선행연구(Baek & Park, 2025; Brown et al., 2020; Park & Choi, 2025)의 논의를 따르면서도, 개발된 문항과 채점 기준의 타당도를 검증하고 다양한 ChatGPT 모델에 따른 차이를 분석하여 그 특징을 도출했다는 점에서 기존 연구와 차별화된다.

둘째, 채점 단계에서는 다수의 자료를 한꺼번에 제시하는 프롬프트 방식보다 맞춤형 GPT를 활용했을 때 채점의 효율성과 일관성이 향상되는 것으로 나타났다. 이는 외부 데이터에 의존하기보다 사용자가 등록한 자료에 기반하여 채점을 수행하면서 채점 기준을 일관되게 적용하기 때문으로 해석할 수 있다. 주목할 점은 맞춤형 GPT에 교사 채점 결과를 추가했을 경우 ChatGPT와 교사 채점 간 일치도가 유의미하게 높아졌다는 것이다. 이러한 결과는 채점 기준과 예시 답안을 제시하는 것으로 한계가 있을 수 있다는 일부 연구(Choi & Park, 2024; Lee et al., 2025)와 맥을 같이하면서도, 교사 채점 결과를 함께 제공했을 때 좀 더 의미 있는 수준의 채점이 가능함을 보여준다. 다만 본 연구에서는 채점에 사용된 사례 수가 적기에 보다 많은 양의 채점 결과를 포함한다면 ChatGPT 채점의 신뢰도를 높일 수 있을 것이다.

이러한 연구 결과가 음악 교과 서·논술형 평가에 시사하는 바는 다음과 같다. 우선, ChatGPT는 수업 내용과 관련지어 새롭고 다양한 문항을 신속하게 생성하면서 음악과 서·논술형 평가의 문항 초안을 작성하는 데에 효과적일 수 있다. 이는 단편적인 음악적

지식이나 사실에 대한 암기력을 측정하거나 감상문 쓰기가 주를 이루어왔던 기존의 서·논술형 평가 방식에서 벗어나 평가의 내용과 범위를 확장하여 분석, 종합, 추론, 응용과 같은 미래 사회에 요구되는 고차원적 능력을 함양할 수 있는 문항을 도출하는 데에 유용한 것으로 보인다. 나아가 최근 강조되고 있는 프로젝트 수업과 관련하여 학습한 내용을 다른 맥락에 어떻게 적용할 수 있는지, 생활 속에서 주어진 문제를 어떻게 해결할 수 있는지 등과 같이 학습의 과정에서 서·논술형 평가가 어떻게 이루어져야 하는지에 대한 시사점을 제공해 준다.

또한, ChatGPT는 채점에 걸리는 시간을 단축하면서 교사에게 부과되는 서·논술형 평가의 채점 업무를 줄이는 데에 도움을 줄 수 있다. 이는 교사당 많게는 10개 이상의 학급을 담당하면서 200명 이상의 학생 답안을 채점해야 하는 중등 음악 교사에게 해당하는 것으로 서·논술형 평가를 지원하는 도구로서 ChatGPT의 활용 가능성을 나타낸다. 하지만 앞서 확인했듯이, 교사 채점 결과를 함께 제공할 때 ChatGPT 채점이 교사와 유사한 수준으로 이루어지기 때문에 평가 전후 단계에서 교사의 전략적인 개입이 요구된다. 다시 말해, 교사는 교육과정 및 수업과 연계하여 평가 계획을 수립하고, 프롬프트를 정교하게 설계 및 수정하며, 자동 채점을 위한 세밀한 채점 기준에 관한 자료를 충분하게 제공하는 한편, 채점 결과를 최종적으로 점검하여 AI가 지니는 한계를 보완할 필요가 있다. 이는 학교 실정, 수업 상황, 학생 수준을 고려하여 평가 맥락에 따라 문항의 적합성을 검토하고, 평가 과제의 특성에 맞게 채점 결과를 분석 및 검증할 수 있는 교사의 비판적 사고와 전문적 판단이 필요함을 시사한다. 더욱이 창의적 표현과 같은 정성적 영역이 중시되는 음악 교과에서 서·논술형 평가의 신뢰성과 타당성을 확보하기 위해서는 평가 설계자이자 최종 검토자로서 음악 교사의 역할이 더욱 중요하다고 볼 수 있다.

이러한 점에서 본 연구는 ChatGPT를 서·논술형 평가의 문항 개발과 채점에 활용하는 과정에서 어떠한 문제가 있는지, 이에 대해 교사가 어떻게 개입해야 하는지를 다각적으로 탐색했다는 점에서 의의를 지닌다. 아울러 서·논술형 평가를 지원하는 도구로 ChatGPT의 가능성과 한계를 함께 확인하였다는 점에서도 의미가 있다. 그러나 본 연구는 채점 과정에서 ChatGPT 모델에 대한 추가 학습이나 내부 파라미터에 대한 조정 없이 프롬프트 엔지니어링과 맞춤형 GPT를 활용하여 결과를 도출했다는 점에서 한계를 지닌다. 또한 ChatGPT와 교사 채점 결과를 비교하는 과정에서 평가 대상의 수가 한정적이며 ChatGPT와 교사 채점 점수 간 상관계수를 분석하는 데에 그쳤다는 점도 본 연구의 제한점이라 볼 수 있다.

이상의 내용을 바탕으로 후속 연구를 제안하면 다음과 같다. 첫째, ChatGPT를 활용하여 평가 문항 개발 및 채점 결과를 도출하기 위한 프롬프트 엔지니어링에 대한 지속적인 연구가 필요하다. ChatGPT에 특정 역할을 부여하거나 채점 기준 및 예시 답안을 제공하

는 것 외에도 평가의 일관성을 높이기 위해 자유도를 최소화하고 문항의 특성에 맞는 정교한 채점이 가능하도록 다양한 프롬프트 방식을 개발 및 적용하는 연구가 꾸준히 이루어질 필요가 있다.

둘째, 사전 학습된 모델에 새로운 데이터를 추가하거나 음악 교과 특성에 적합한 독자적인 생성형 AI 평가 모델을 구축하는 연구가 필요하다. 아직 ChatGPT에 음악 관련 데이터가 풍부하지 않다는 점을 고려했을 때, 더 많은 학생 답안과 교사 채점 결과를 포함하여 모델을 정교화한 후에 채점 결과를 비교·분석할 필요가 있다. 이는 ChatGPT와 교사 간 채점 일치도를 향상할 수 있을 뿐만 아니라 학생들에게 유용하고 개별화된 피드백을 제공할 수 있을 것이다.

셋째, 생성형 AI 채점의 정확성과 일관성을 높이기 위해 음악 교사의 평가 전문성을 신장하기 위한 연구도 병행될 필요가 있다. 전술하였듯이 ChatGPT 단독으로 채점을 수행하기에는 무리가 있으며 전략적으로 교사가 개입할 필요가 있었다. 이를 위해서는 국가 혹은 교육청 차원에서 서·논술형 평가를 위한 평가 문항과 채점 기준을 개발 및 보급하고 평가 관련 체계적인 연수를 통해 음악 교사의 평가 역량을 키우기 위한 노력이 수반되어야 한다. 이와 더불어, 음악 교사의 생성형 AI를 활용한 평가 경험을 구체적으로 탐색함으로써 실제 현장에서 어떠한 부분에서 도움이 되는지 혹은 도움이 되지 않는지, 학생들에게는 어떠한 긍정적 혹은 부정적 영향을 주는지 그 효과와 한계를 심층적으로 탐색하는 실험연구도 병행할 필요가 있다. 추후 음악 교육 분야에서 서·논술형 평가에 관한 관심과 논의, 음악 교사와 생성형 AI와의 협력을 통해 서·논술형 평가의 활성화 방안을 모색하는 연구가 이어지길 기대한다.

References

- Baek, M., & Park, C. (2025). A study on preservice home economics teachers' experiences in developing essay-type assessment items using generative AI. *Comprehensive Research in Education*, 23(1), 153-168. <https://doi.org/10.31352/JER.23.1.153>
- Baek, S. (2002). *Performance assessment: Theoretical perspectives*. Kyoyook Kwahaksa.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. <https://doi.org/10.48550/arxiv.2005.14165>

- Choi, S., & Park, J. (2024). Analysis of in-service Korean-language teachers' development of essay-type assessment items using generative AI. *Journal of CheongRam Korean Language Education*, 97, 243-270. <https://doi.org/10.26589/jockle..97.202401.243>
- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Lawrence Erlbaum Associates.
- García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences*, 15(8), 946. <https://doi.org/10.3390/educsci15080946>
- Gyeonggido Office of Education (2025). *Guidebook for essay-type assessment in secondary education*. Gyeonggido Office of Education.
- Ham, E., Park, S., Lee, B., Kim, G., & Lee, D. (2024). Protocol for developing constructed-response items using ChatGPT and quality evaluation: Focusing on Korean language arts. *Korean Journal of Educational Research*, 62(8), 63-93. <https://doi.org/10.30916/KERA.62.8.63>
- Han, S., & Han, K. (2025). Effects of parameter-based AI composition on self-efficacy and interest in elementary music education. *Korean Journal of Research in Music Education*, 54(3), 283-304. <https://doi.org/10.30775/KMES.54.3.283>
- Holster, J. (2024). Augmenting music education through AI: Practical applications of ChatGPT. *Music Educators Journal*, 110(4), 36-42. <https://doi.org/10.1177/00274321241255938>
- Im, J., & Lim, J. (2024). Action research on the development program of ChatGPT for automated scoring of essay type items. *Journal of Educational Innovation Research*, 34(1), 349-370. <https://doi.org/10.21024/pnuedi.34.1.202403.349>
- Jeong, Y., & Baek, S. (2025). Comparison of the usefulness of prompt-engineering techniques for automatic scoring of social-studies essay-type items. *Korean Journal of Educational Research*, 63(4), 95-124. <https://doi.org/10.30916/KERA.63.4.95>
- Ji, M., Oh, H., Noh, M., Kwon, U., Kim, Y., Lee, J., Jang, H., So, M., & Cho, B. (2024). *Evaluating the 2022 revised curriculum: Adding wings with AI*. AnswerBook.
- Kim, H. (2024). Educational significance and application directions of descriptive and essay-type assessment in the 2022 revised art curriculum. *Art Education Review*, 92, 27-44. <https://doi.org/10.25297/AER.2024.92.27>
- Kim, K. (2020). Exploring the evaluative meaning of essay-type assessment. *Journal of Educational Evaluation Research*, 33(4), 839-862. <https://doi.org/10.31158/JEEV.2020.33.4.839>
- Kim, S., & Hong, J. (2024). Popular music education using generative artificial intelligence:

- Focusing on ChatGPT. *Korean Journal of Popular Music*, 33, 35-68. <https://doi.org/10.36775/kjpm.2024.33.35>
- Kwon, T. (2024). A study on writing scoring and feedback practices using ChatGPT: Focusing on prompting strategies. *The Academy for Korean Language Education*, 141, 7-42. <https://doi.org/10.15734/koed..141.202412.7>
- Lee, B. (2015). *Development and application of essay-type assessment items in middle school music education*. Master's thesis, Korea National University of Education.
- Lee, J., Lee, E., & Kim, K. (2025). Enhancing alignment between large language models and teacher in open-ended assessment through in-context Learning. *Brain, Digital, & Learning*, 15(3), 375-401. <https://doi.org/10.31216/BDL.2025.15.3.4>
- McMillan, J. H. (2013). *Classroom assessment: Principles and practices for effective standards-based instruction* (S. Son, J. Park, S. Kang, C. Park, & K. Kim, Trans.). Kyoyook Kwahaksa.
- Ministry of Education (2022) *Music curriculum*. No. 2022-33. Ministry of Education.
- Ministry of Education (2023, October 10). *Draft plan for the 2028 university admission system reform*. Retrieved June 10, 2025, from <https://moe.go.kr>
- Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. *Paper presented at the 17th European Conference on Technology Enhanced Learning* (pp. 243-257), Toulouse.
- OpenAI (2025). *ChatGPT-5*. Retrieved August 10, 2025, from <https://platform.openai.com/docs/models>
- Park, G., & Choi, S. (2025). A study on automatic item generation for multiple-choice assessment in the reading domain of Korean language education. *Journal of Curriculum and Evaluation*, 28(1), 215-246. <https://doi.org/10.29221/jce.2025.28.1.215>
- Park, H., Kim, S., Kim, K., Lee, M., & Kim, K. (2019). *Plans for enhancing the quality of essay-type assessment in individual schools*. Research Report ORM 2019-54-10. Korea Institute for Curriculum and Evaluation.
- Park, J. (2024). An exploration of the current introduction and implementation issues of essay-type assessment in Korean language education. *Journal of Cheong Ram Korean Language Education*, 101, 273-307. <https://doi.org/10.26589/jockle..101.202409.273>
- Park, S., Hong, Y., & Lee, B. (2024). A study on exploring the potential of ChatGPT in writing skills assessment: Focusing on essay writing. *Korean Journal of Educational Research*, 62(5), 219-248. <https://doi.org/10.30916/KERA.62.5.219>
- Park, S., Lee, B., Ham, E., Lee, Y., & Lee, S. (2023). Exploring the possibility of science-inquiry

- competence assessment by ChatGPT-4: Comparisons with human evaluators. *Korean Journal of Educational Research*, 61(4), 299-332. <http://dx.doi.org/10.30916/KERA.61.4.299>
- Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: A systematic review and applications. *arXiv*. <https://doi.org/10.48550/arxiv.2405.11029>
- Seong, J., & Shin, B. (2023). Exploring the feasibility of automatic scoring of written tests using ChatGPT: Focusing on the world geography written test. *Journal of the Association of Korean Geographers*, 12(3), 415-432. <https://doi.org/10.25202/JAKG.12.3.3>
- Seong, T., Lim, H., Jeon, K., & Choi, Y. (2025). *Foundations of educational evaluation* (4th ed.). Hakjisa.
- Shaw, B. P. (2024). Artificial intelligence and assessment: Three implications for music educators. *Music Educators Journal*, 111(2), 19-25. <https://doi.org/10.1177/00274321241296118>
- Shin, B., Lee, J., & Yoo, Y. (2024). Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. *The Mathematical Education*, 63(2), 187-207. <https://doi.org/10.7468/mathedu.2024.63.2.187>
- Song, J. (2022). *Improvement plans for essay-based assessment and university admissions in the transition to a future-oriented education system*. Daegu Education 2022-102. Daegu Metropolitan Office of Education.
- Song, M., Kim, D., Kim, J., Park, S., Park, J., & Jeong, S. (2024). *A study on the application of AI models for automated grading of subject-specific essay-type assessments*. RRE 2024-9. Korea Institute for Curriculum and Evaluation.
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education. Artificial Intelligence*, 3, 100075. <https://doi.org/10.1016/j.caeai.2022.100075>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. <https://doi.org/10.48550/arxiv.2201.11903>
- Yoon, K. (2024). A study on teaching plans for music creation applying generative AI based on 'Text to Music' to group investigation model. *Korean Journal of Research in Music Education*, 53(1), 143-164. <https://doi.org/10.30775/KMES.53.1.143>