Korean Journal of English Language and Linguistics, Vol 23, November 2023, pp. 1077-1093
DOI: 10.15738/kjell.23..202311.1077

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

대학생 영어 능력 평가를 위한 수준설정 사례 연구

이용상 (인하대학교) · 김은주 (한양여자대학교)

Received: October 18, 2023
Revised: November 4, 2023
Accepted: November 15, 2023

Yongsang Lee
(First author)
Associate Professor, Dept. of Education, Inha University
Tel: 032-860-7873
E-mail: yong21c@gmail.com

Eun-Ju Kim
(corresponding author)
Associate Professor, Dept. of English, Hanyang Women's University
Tel: 02-2290-2280
E-mail: exk188@hywoman.ac.kr

ABSTRACT

Yongsang, Lee and Eun-Ju Kim. 2023. A case study on standard setting for English proficiency test for university students. *Korean Journal of English Language and Linguistics* 23, 1078-1094.

This study applied a standard setting method for determining cutoff scores for university students' English proficiency tests, which are criterion-referenced tests, conducted at universities and explored the application of this approach for enhancing the effectiveness of general English education at different levels. To differentiate students' proficiency levels, the study used the cutoff scores determined by the Angoff method based on the English proficiency test items used for sub-divisions of general English courses at University A. University A conducts its sub-division classes in three proficiency levels, and, accordingly, this study employed level determination based on the three performance level descriptions(PLD). The results of the Angoff standard setting indicated that, as rounds of discussion progressed, the panel members tended to make decisions in favor of raising the cutoff scores, and the standard deviation of the cutoff scores estimated by the panel members tended to decrease demonstrating a narrowing of opinions among the panel members. However, it was confirmed that the estimation tendencies of specific panel members could influence the overall standard deviation, and, therefore, guidelines are needed to control such bias in future standard settings. The application of the cutoff scores determined through standard setting to assess students' English proficiency levels showed a difference in the distribution of level-based classes from what is currently being conducted at University A. This suggests that, for universities using fixed cutoff scores to divide classes into proficiency levels, the standard setting is necessary to ensure the appropriate division of level-based classes.

KEYWORDS

Standard setting, criterion-referenced test, university student English proficiency test, performance level description.

1. 서론

국내 대학들은 신입생들을 우선 대상으로 교양영어, 대학영어, 혹은 기초영어라는 이름으로 다양한 영어 강좌를 교양교과로 제공하고 있다. 각 대학들의 개별 상황에 따라 추구하는 교육목표, 제공되는 강좌 수, 강좌 종류 그리고 운영방식은 다양하나, 프로그램 운영 목표나 내용 변화의 방향성에는 공통점이 있다고 할 수 있다.

1990년대 초반까지 국내 대학들이 제공하던 교양영어 강좌들은 말그대로 학생들의 교양 지식 습득 및 인문 소양 함양을 목표로 대부분 영어 텍스트 독해 활동 위주, 교수자 중심 형태로 진행되었다(오은주 2022). 그러나 실용 영어능력 함양에 대한 사회와 학습자들의 요구가 높아짐에 따라, 2000년대에 교양영어의 목표는 말하기와 쓰기 활동 등을 포함한 표현 능력 기반 영어 능숙도 증진으로 변화되어 왔다(김성혜, 임자연 2013, 김성해 외 2021).

이러한 실용영어능력 함양을 목표로 하는 교양영어 수업 운영 중 교수자나 학습자들이 경험한 공통적 어려움은 학습자들 간의 수준차이다(정양수 2014). 이러한 수준차로 인하여, 교수자 입장에서는 지도에 적절한 수준 파악이 어렵고, 학습자 입장에서는 수업의 수준이 적절한 경우보다 너무 쉽거나 너무 어렵다는 불만이 제기되었다. 그 대안으로 외국어 교육에서 활용되는 수준별 수업이 교양영어 수업에도 점짐적으로 도입되어 운영되고 있는 실정이다. 수준별 수업에 대한 학습자들의 만족도는 대체로 높은 것으로 나타났고(서정아 2018, 서혜진 2020), 특히 낮은 수준의 학습자들이 느끼는 만족도가 높은 것으로 보고되었다(김성해 외 2021).

수준별 교양 영어 수업 운영을 하는 대학들은 진단평가를 통하여 수준별 수업을 위한 학생들의 영어 능숙도 수준을 진단하여 구분해야 한다. 따라서, 대학들은 자체 진단평가를 개발하여 실시하거나, TEPS나 TOEIC, 혹은 수능외국어영역 점수를 진단평가에 활용하고 있다(김우형 2015). 한편, 이러한 진단평가를 실시하는 대부분의 대학들은 학생의 수준을 구분하는 데 있어서 검사지의 난이도와 교양영어에서 상정하는 학생들의 수준에 대한 기준을 고려하지 않고 학생들의 등급을 분할하는 고정분할점수 방식을 이용하여 학생들의 수준을 나누고 있다. 이와 같은 방법은 학생들의 수준을 구분하는 데 편리하지만, 검사지의 난이도 변화나 응시자 집단의 영어 능숙도 차이를 고려하지 못하므로 학생들의 영어 능숙도를 일관성 있게 구분할 수 없다는 단점이 있다. 이와 같은 단점으로 인해 고정분할점수를 이용할 경우에는 학생들의 능숙도 수준을 정확히 파악하여 수준별 수업을 구현하려는 목적을 달성할 수 없다는 문제점을 가지고 있다.

이러한 문제점으로 인해 절대평가로 알려져 있는 준거참조평가는 수준설정이라는 별도의 절차를 통해 등급 또는 수준을 구분하는 기준(준거)를 정하고 있다. 과거 준거참조평가로 개발되었던 국가영어능력평가시험이나 현재 국가단위로 학생들의 학업 성취도를 점검하는 국가수준학업성취도평가 등에서는 모두 수준설정 절차를 통해 학생들의 영어 능숙도 또는 학업 성취도 수준을 구분하는 기준인 분할점수를 산출하고 있다. 그러나 상술한 바와 같이 아직까지 국내 대부분의 대학에서는 학생들의 영어 능숙도를 준거참조평가로 평가하고 있음에도 불구하고 수준설정의 절차를 통해 학생들의 영어 능숙도 수준을 구분하기 위한 분할점수를 산출하여 적용하는 경우는 찾아보기 힘든 실정이며 이와 관련한 연구보고도 부족한 상황이다.

따라서, 본 연구에서는 대학에서 학생들의 영어 능숙도를 평가하는 진단평가에 분할점수 산출을 위한 수준설정 방법을 적용해 봄으로써 대학의 진단평가에서 안정적인 수준설정 및 분할점수

산출 과정에서의 문제점과 이슈를 점검해 보고 그 결과를 토대로 대학의 영어 능숙도 평가에서 내실 있는 준거참조평가가 이루어질 수 있도록 하기 위한 시사점을 도출하고자 한다. 이를 위해 본 연구에서는 A대학의 교양영어 수준별 분반을 구성하기 위해 학생들의 영어 능숙도를 구분하기 위한 기준을 도출하고자 수준설정을 시행하였다. 수준설정에는 A대학에서 활용한 영어 진단평가 문항과 채점 자료를 이용하였다.

2. 이론적 배경

2.1 수준설정

수준설정은 준거참조평가(criterion-referenced test)에서 피험자들의 능력 수준을 구분하는 분할점수(cut-off score)를 결정하는 절차이다. 수준설정은 준거참조평가에서 학생들의 절대적인 능력 수준을 기술한 성취수준기술(PLD)을 기준으로 피험자들이 응시한 시험의 난이도 등을 고려하여 피험자들의 성취수준을 나눌 수 있는 분할점수를 내용전문가들(이하 패널)이 숙의 과정을 통해 결정하는 과정이다. 이와 같은 수준설정의 대표적인 방법으로는 원점수를 기준으로 분할점수를 산출하는 앵고프 방법(Angoff method)과 문항반응모형 분석을 통해 산출되는 능력점수를 기준으로 분할점수를 산출하는 북마크 방법(bookmark method) 등이 있다.

가장 일반적으로 사용되는 앵고프 방법은 PLD에 수준별 피험자들이 갖추어야할 능력, 기술, 태도 등을 최소한으로만 갖춘 피험자를 상정하고 이러한 피험자가 개별 문항에 정답을 맞힐 확률을 추정하여 분할점수를 산출하는 방식이다. 수준설정에서는 수준별로 최소한의 능력만을 갖춘 피험자를 최소능력자(minimum competency person; MCP)라 지칭하며, 앵고프 방법에서는 일련의 패널들이 PLD를 기준으로 MCP에 대한 논의를 거쳐 구체화하는 작업을 진행하여 MCP에 대한 합의를 도출한다. 앵고프 방법에서는 이렇게 패널 간 합의된 MCP가 개별 문항에서 정답을 맞힐 확률이 어느 정도인지 개별 패널들이 추정하도록 한 후, 추정된 값의 중앙값을 등급 분할점수로 정하게 된다. 초기의 앵고프 방법은 패널들이 정답률 추정을 1회만 실시하므로 패널 간 합의를 이루는 절차가 생략되어 있었으며, 이러한 제한점을 보완하여 수정된 앵고프 방법은 최소 3회 이상의 정답률 추정 과정을 통해 패널 간 충분한 토의와 합의가 이루어지도록 하고 있다. 최근에는 수정된 앵고프 방법이 널리 사용되며 따라서 일반적으로 앵고프 방법은 수정된 앵고프 방식을 의미한다.

앵고프 방법과 비교하여 북마크 방법은 문항반응모형 분석을 통해 개별 문항의 난이도를 추정하고 이를 기준으로 우선 쉬운 문항부터 어려운 문항 순서로 배열된 문항집(ordered item booklet; OIB)을 만든다. 북마크 방법에서는 이 순서화된 문항집을 사용하여 MCP가 맞출 수 있는 마지막 문항에 표시(bookmarking)하는 방식으로 등급 분할 문항을 식별하고, 이 등급 분할 문항에 매칭되는 능력점수를 개별 패널이 추정한 등급 분할점수로 간주하며, 앵고프 방법과 마찬가지로 추정된 값의 중앙값을 등급 분할점수로 정한다.

이와 같이 앵고프 및 북마크 방법은 문항의 내용에 기반한 방법으로, 개별 문항에 대한 분석결과를 바탕으로 수준설정을 실시하는 방법으로 선다형 문항에 적합한 방법이다. 이와

비교하여 서답형 문항은 앵고프 방법과 패널이 개별 문항에 대한 정답률을 추정하는 데 어려움이 있고, 비교적 적은 수의 문항으로 검사지가 구성되기 때문에 북마크 방법과 같이 OIB를 제작하여 활용하는 데 한계가 있다. 따라서 이러한 서답형 문항에 대한 수준설정은 문항보다 피험자의 응답에 초점을 맞추는 수행 프로파일 방법(performance profile approach, Zieky et al. 2008)이 최근 많이 활용되고 있다. 수행 프로파일 방법은 앵고프나 북마크 방법과 달리 피험자의 응답에 초점을 맞추어 수준을 설정하는 방법이며, 토익의 쓰기 시험 등에서(Educational Testing Service 2021)에서 활용된 바 있다.

2.2 초등영어교육 연구 동향 관련 선행연구

영어 검사에서의 수준설정에 대한 선행 연구들을 살펴보면 수준설정 방법 비교에 초점을 맞춘 연구들(곽예린, 김경리 2022, 이선경, 지은림 2017, 임의진, 이용원, 전희성 2019)과 수준설정 방법론에 대한 연구들(박인용, 송미영, 김성숙 2014, 박연복, 이규민, 강상진 2011), 그리고 수준설정 실행 연구들(김성숙, 박은아, 서민희 2014, 이영주 외 2017)로 나누어 볼 수 있다.
우선 곽예린, 김경리(2022)는 앵고프 방법과 북마크 방법을 비교하기 위해 종합진로직업적성검사 중학생용 데이터(성태제 외, 2016)를 이용하여 두가지 방법으로 수준설정을 실시하고 그 결과를 비교하였으며, 임의진 외(2019)는 TEPS의 성취수준을 유럽 언어 공통 기준(common European framework of reference for languages; CEFR)에 연계함에 있어 앵고프와 북마크 방법을 비교하였다. 또한 이선경과 지은림(2017)은 중학생용 인성 검사 자료를 활용하여 라쉬 방법(Rasch method)과 앵고프 방법을 비교하였으며 수준설정 결과에 있어서 두 방법간 상당히 높은 수준에서의 일관성이 있음을 확인하였다.

한편 수준설정 방법론에 대한 연구로서 박인용 외(2014)는 앵고프 방법을 이용한 수준설정에서 정답률 추정의 타당성을 검토하였으며, 이를 통해 패널들이 문항이 쉬운 경우에는 정답률을 과소추정하거나 문항이 어려울 경우에는 정답률을 과대 추정하는 경향이 있음을 확인하였다. 박연복 외(2011)에서는 군집분석을 통한 수준설정 방안을 제시하고, 이를 통해 산출된 분할점수와 기존의 북마크 방법을 이용해 산출한 분할점수 간의 비교를 통해 군집분석을 통한 수준설정 방안의 타당성을 검토하고, 그 활용 가능성을 확인하였다.

이외에 분할점수 산출을 위한 수준설정 연구로서 김성숙 외(2014)는 준거참조평가인 성취평가제를 위한 분할점수 산출 방안으로 고정분할점수와 앵고프 및 에벨(Ebel) 방법을 적용한 결과를 비교하고 이들 방법을 성취평가제에서 활용할 경우 예상되는 문제점들을 검토하였다. 이영주 외(2017)는 의대의 준거참조평가에 앵고프 방법과 에벨 방법을 적용하여 분할점수를 산출해보고 그 결과를 비교 분석함으로써 의대 수업에서 준거참조평가와 이를 위한 분할점수 산출 방법들의 활용 가능성을 탐색하였다.

이상의 연구들을 살펴보면, 수준설정과 관련된 선행 연구들은 수준설정 방법의 비교나 수준설정 방법의 타당성 검토와 같이 수준설정 방법론 자체에 초점을 맞춘 연구들이 주를 이루었음을 알 수 있다. 특히, 영어 시험을 준거참조평가로 시행하고 학생들의 성취수준을 구분하기 위한 분할점수를 산출하기 위해 수준설정 방법을 활용하는 활용연구는 매우 제한적이었으며, 따라서 본 연구에서는 영어 검사에서 학생들의 성취수준을 구분하기 위해 신뢰롭고 타당한 분할점수를

산출하기 위한 노력으로써 A대학 영어 진단 검사에서의 수준설정 사례를 분석하여, 대학 영어 능숙도 검사에서 학생들의 성취수준을 구분하기 위한 분할점수를 안정적이고 일관성 있게 산출하기 위한 시사점을 도출하였다.

3. 연구 방법

3.1 연구 대상

본 연구의 피험자들은 22년도 2학기에 A대학의 교양영어 수강자 1,698명 중, 학기말에 진행된 진단검사 중, B세트 진단검사에 응시한 354명이다. 진단검사는 100점 만점 두 세트(A, B세트)로 나누어 진행되었다. 피험자들은 수강 중인 수업 수준이 고르게 분포된 두 집단으로 나누어져 진단검사에 응시하였다. 진단검사는 학교 학습관리시스템(learning management system: LMS)를 활용하여 온라인으로 진행되었으며 응시 시간은 총 40분이였다. 354명의 응시자는 22학번이 320명, 21학번이 28명, 20학번이 3명, 19학번이 2명, 마지막으로 18학번 1명으로 구성되었으며, 총 15개 학과 재학생들이 응시하였다.

3.2 대학 영어 진단 검사 및 PLD

본 연구에 활용된 평가지는 A대학의 교양영어 수강 대상자의 수준별 분반을 위하여 사용되는 진단 검사지이다. 진단평가는 총 35문항으로 이루어져 있고, 이 문항들은 문항별 배점을 차별화하여 100점 만점 검사지로 구성되어 운영된다.

A대학의 교양영어 교육의 목적은 실용영어능력 향상에 중점을 두고 있기에 진단평가 영역에 따른 문항분포는 회화(9문항), 어휘(10문항), 문법(7문항), 독해(9문항)로 이루어져 있고 평가내용은 주로 생활 영어 위주 주제를 다루고 있다(그림 1 문항 예시). 문항별 배점은 난이도에 따라 2점~4점 사이로 분포되어 있다.

<취미에 관한 대화>

A: What kind of movies do you like?

B: 글쎄요, 전 그다지 영화를 좋아하지 않아요.

(A) Well, I can't stand the movie.

(B) Well, I'm not really a movie fan.

(C) Well, I prefer seeing a movie.

(D) Well, what about you?

그림 1. 문항 예시

본 연구에 활용된 대학 영어 진단 검사는 8개의 1지문 2문항을 포함하여 총 35개의 선다형 문항으로 구성되어 있으며, 문항별 정답률은 그림 2와 같다.

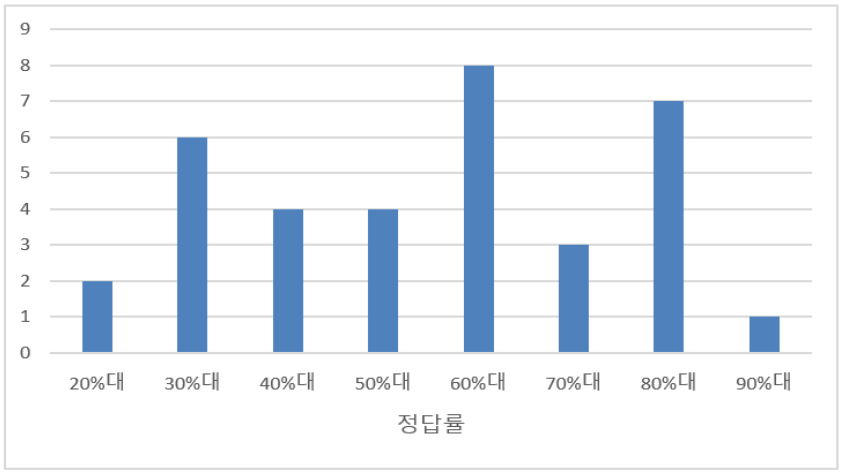

그림 2. 정답률 구간별 문항 수

수준설정을 위해서는 우선 개별 성취수준을 구분하기 위한 기준이라 할 수 있는 성취수준기술(PLD)이 필요하다. PLD는 개별 성취수준에 해당하는 학생들이 가지고 있는 능력과 기술 등에 대한 기술(description)로서 해당 수준의 학생이 무엇을 할 수 있고 무엇을 할 수 없는지에 대한 기술을 포함한다. 수준설정은 이러한 PLD를 먼저 기술하고 확정하는 작업으로부터 시작한다고 할 수 있으며, 본 연구에서는 수준설정 작업을 위한 PLD를 다음과 같은 절차를 통해 제작 및 확정하였다. 먼저, 평가 전문가와의 본 대학 교양영어 수업의 내용, 진행 방식, 평가 등에 면담을 통하여 읽기 영역 PLD만이 현 상황에서 가장 적절하다는 자문을 받았다. 그 이유는 본 교양영어 강좌 수강 후 기대되는 학습자들의 학업성취도를 지필평가 형태로 진행한다는 점과 평가의 내용이 읽기/독해에 기반한 이해도 측정이 주를 이루고 있다는 점이다. 이후, 읽기 영역 PLD 참고자료로 국가영어능력평가시험(national English ability test; NEAT)의 읽기 3등급 PLD와, 미국 외국어 교육협회(American council on the teaching of foreign languages: ACTFL)의 읽기 영역 능숙도 가이드라인을 바탕으로 현재 A대학에서 각기 다른 수준(상, 중, 하)의 교양영어를 가르치는 3인의 교수자와 연구자가 함께 본 대학의 교양영어 PLD 제작을 완료하였다(표 1. 교양영어 PLD).

표 1. 교양영어 PLD

수준	읽기
Novice-Mid (NM)	- 실생활의 친숙한 주제에 대한 글의 중심내용 및 세부내용을 제한적으로 이해(또는 추론)한다. - 그림, 도표와 같은 시각자료와 함께 제시된 글의 정보를 제한적으로 이해한다. - 간단한 글의 논리적 전개와 문장구조 등을 제한적으로 이해한다. - 필자의 입장이나 글의 흐름을 제한적으로 이해한다.
Novice-High (NH)	- 실생활의 친숙한 주제에 대한 글의 중심 내용 및 세부 내용을 대체로 이해(또는 추론)한다. - 그림, 도표와 같은 시각 자료와 함께 제시된 글의 정보를 대체로 이해한다. - 비교적 복잡한 글의 논리적 전개와 문장 구조 등을 대체로 이해한다. - 필자의 입장이나 글의 흐름을 대체로 이해한다.
Intermediate-low (IL)	- 실생활의 친숙한 주제에 대한 글의 중심 내용 및 세부 내용을 정확하게 이해(또는 추론)한다. - 그림, 도표와 같은 시각자료와 함께 제시된 글의 정보를 정확하게 이해한다. - 비교적 복잡한 글의 논리적 전개와 문장구조 등을 정확하게 이해한다. - 필자의 입장이나 글의 흐름을 정확하게 이해한다.

3.2 수준설정 방법: 앵고프 방법

본 연구에서는 선다형 문항으로 구성되어 있는 검사지의 특성과 원점수 기반의 점수 체계를 고려하여, 원점수로 분할점수를 산출할 수 있는 앵고프 방법으로 수준설정을 진행하였다. 앵고프 방법은 분할점수 추정 절차가 1회만 시행이 되므로 패널 간 편차 조정이 이루어질 수 있는 기회가 없는 문제점이 있다. 이러한 문제점을 극복하기 위해 분할점수 추정 절차를 3차례 반복하는 수정된 앵고프 방법이 일반적으로 사용되고 있으며, 최근에는 이러한 수정된 앵고프 방법을 앵고프 방법으로 지칭하고 있다. 본 연구에서는 총 8명의 내용전문가 패널이 참여하였으며, 8명의 패널은 A대학 교양영어 강의를 담당하는 교·강사들이다. 이들 패널은 전원 영어 교육 관련 분야 박사학위를 소지하고 있고, 개인별로 7명은 A대학에서 교양영어를 가르친 경력이 3년~5년 사이이고, 나머지 1명은 교양영어 강의를 직접 진행한 경험은 수준설정 당시에는 없었다. 본 연구에서 진행한 앵고프 수준설정의 절차를 요약하면 다음과 같다.

① 영어 시험의 평가틀과 성취수준의 특성에 대하여 설명한다. 이후 패널들은 'Novice-Mid(NM)', 'Novice-High(NH)', Intermediate-low(IL)'의 세 가지 성취수준의 개념과 각 성취수준의 최소 능력에 대한 개념을 토의하고 패널들 간의 합의점을 도출한다.
② 패널들은 연습용 문항에 대하여 각 성취수준의 MCP 집단별 정답률을 예측해 봄으로써, 수준설정 절차, 성취수준 구분 기준 및 성취수준에 대해 공감대를 형성한다.
③ 패널들에게 검사지를 나누어 주고, 전체 문항의 내용, 정답을 면밀히 검토하도록 한다.
④ 패널들은 전체 문항에 대하여 2개의 성취수준(IL, NH) 각각에서 MCP 학생들의 기대 정답률을 판정한다(1라운드).
⑤ 패널들은 1라운드에서 각자가 판정한 기대 정답률과 전체의 기대 정답률 중앙값 및 최댓

값, 최솟값을 검토한다. 패널 간 기대 정답률의 편차가 큰 문항부터 논의하여 해당 문항의 난이도에 대한 개념적 수준에서 합의를 도출한다.

⑥ 패널들은 2개의 성취수준(IL, NH) 각각에서 MCP 학생들의 기대 정답률을 판정한다(2라운드)

⑦ 2라운드 결과를 바탕으로 각 성취수준에서의 두 번째 분할점수를 산출하고, 두 번째 분할점수를 기준으로 성취수준별 비율을 산출한다.

⑧ 2라운드에서 각자가 판정한 기대 정답률과 전체의 기대 정답률 중앙값 및 최댓값, 최솟값을 검토한다.

⑨ 전체 문항에 대하여 2개의 성취수준(IL, NH) 각각에서 MCP 학생들의 기대 정답률을 판정한 후, 각 성취수준에서의 세 번째 분할점수를 산출한다(3라운드).

⑩ 세 번째 분할점수를 기준으로 성취수준별 비율을 산출하고 그 결과를 참고하여 수준설정 과정의 종료 여부를 결정한다.

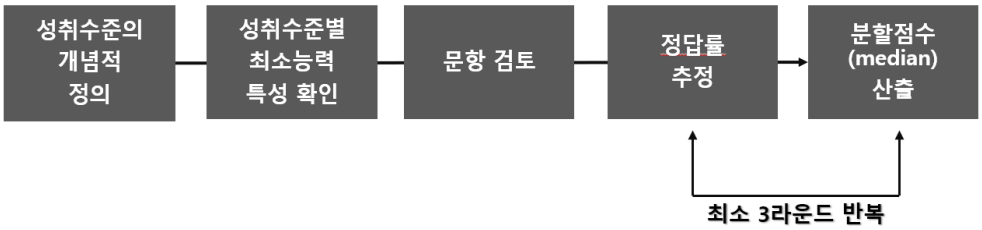

그림 3. 앵고프 수준설정 절차

본 연구에서는 총 3라운드에 걸쳐 수준설정을 진행하였으며, 3라운드에서 도출된 분할점수를 영어 진단평가를 통해 구분할 수 있는 학생들의 성취수준 분할점수로 확정하였다.

4. 연구 결과

본 연구에서는 8명의 패널들이 합의한 등급별 MCP를 기준으로 등급별 MCP들이 개별 문항에 대해 정답할 확률을 우선 추정하였다. 본 연구의 영어 검사에서는 3개의 성취수준(IL, NH, NM)으로 구분하므로, 이들 성취수준을 구분하기 위해서는 총 2개의 분할점수가 필요하다. 이를 위해 가장 높은 성취수준인 IL의 MCP와 그 다음 성취수준인 NH의 MCP를 상정하여 이들 MCP의 문항별 정답률을 추정하는 방식으로 2개의 분할점수를 산출하였다. 분할점수는 패널 간 편차와 산출된 분할점수를 적용하였을 때의 수준별 인원 비율을 고려하여 최종 확정하였으며, 이를 위해 총 3라운드에 걸쳐 문항별 정답률 추정을 실시하였다. 이후 추정된 정답률을 문항당 배점에 곱하여 기대 점수를 계산하였고 이들 더하여 최종 원점수 기준에서의 등급 분할점수를 산출하였다. 영어 검사의 문항당 배점은 2점에서 4점 사이이며, 35문항의 총점은 102점 만점으로 분할점수는 102점 만점을 기준으로 산출하였다.

우선 IL과 NH 분할점수와 NH와 NM 분할점수를 산출하기 위해 IL과 NH의 MCP의 문항별 정답률 추정 결과를 살펴보면, IL 수준의 MCP는 60%~90%의 정답률을 보일 것으로 예측되었으며, NH의 MCP는 45%~75%의 정답률을 보일 것으로 예측되었다. 이와 같은 문항별 정답률은 그림 4에서 보여주듯이 IL 수준과 NH 수준의 MCP가 분명히 구분되고 있음을 확인할 수 있었다. 이와 같은 결과는 패널들이 상정하는 IL 수준과 NH 수준의 MCP가 선명하게 구분되는 수준의 MCP임을 보여주는 것으로 본 연구에서 사용된 PLD에서 제시하는 성취수준이 분명히 구분될 수 있는 것임을 실증적으로 보여준다.

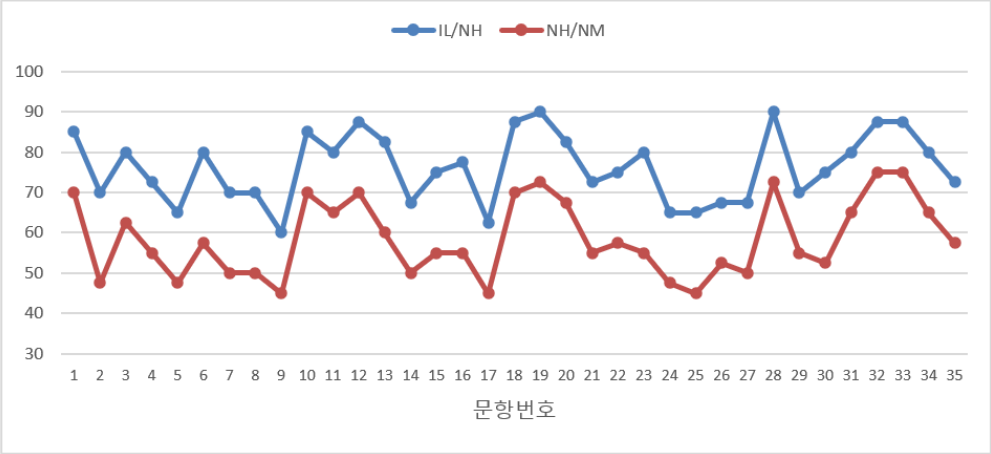

그림 4. 문항별/등급별 MCP 정답률 추정 결과(최종)

이는 표 2의 패널별 분할점수 산출 결과에서도 확인할 수 있었다. 요컨대, 패널별로 산출한 성취수준 분할점수의 경우, IL과 NH 분할점수는 63점~ 82점이었으며, NH와 NM 분할점수는 47점~68점으로, 패널별로 두 분할점수의 차이를 적게는 12점에서 많게는 23점으로 차이를 두고 있어, 영어 검사에서 상정하는 3개의 성취수준이 실증적으로 피험자에게서 구분될 수 있는 성취수준임을 확인할 수 있었다. 더욱이 그림 5에서 보여주듯이 대부분의 패널이 추정한 NH/NM 분할점수가 IL/NH 분할점수들 보다 높지 않아 이들 패널 간에 3개의 수준을 구분하기 위한 MCP의 설정에 충분한 공감대가 형성되었음을 확인할 수 있었다.

표 2. 패널별 성취수준 분할점수 및 분할점수 차이(최종)

패널	IL/NH	NH/NM	분할점수 차이
A	79.2	55.8	23.5
B	66.2	50.8	15.5
C	72.4	53.4	19.0
D	63.1	46.5	16.6
E	82.1	68.0	14.1
F	80.4	66.8	13.6
G	71.7	59.7	12.0

H	79.2	58.0	21.2
최대	82.1	68.0	23.5
최소	63.1	46.5	12.0

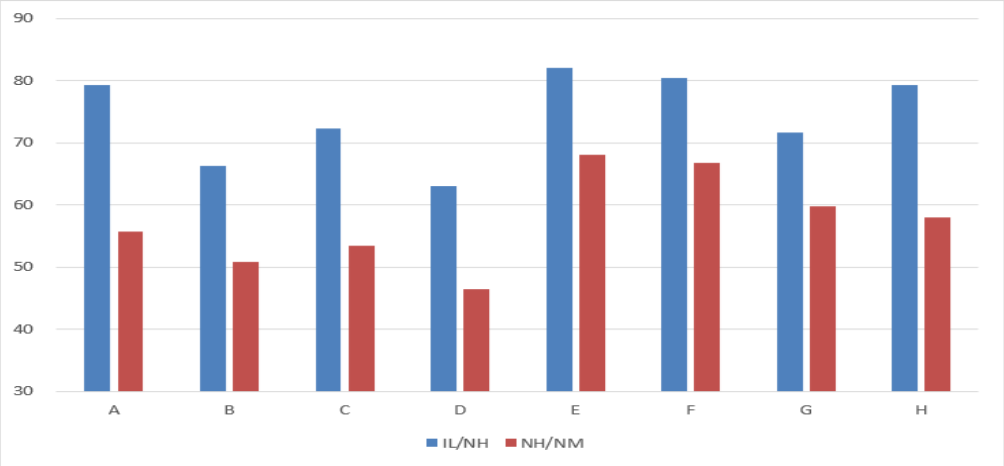

그림 5. 패널별 분할점수 산출 결과(최종)

다음으로 수준설정 결과를 살펴보면, IL과 NH 수준을 구분하는 분할점수와 NH와 NM 수준을 구분하는 분할점수를 총 3회에 걸친 앵고프 수준설정을 통해 산출한 결과 표 3과 그림 6에서 보여주듯이 라운드를 거듭할수록 분할점수가 높아지는 경향을 확인할 수 있었다. 예컨대 IL과 NH 수준을 구분하는 분할점수는 1라운드에서 61점으로 산출되었으나 2라운드에서 73.2, 3라운드에서는 75.8로 높아졌고, NH와 NM 수준을 구분하는 분할점수는 1라운드에서 43점, 2라운드에서 52.6, 3라운드에서 56.9로 높아졌다.

표 3. 라운드별 분할점수 및 표준편차

라운드	IL/NH		NH/NM	
	분할점수	표준편차	분할점수	표준편차
1R	61.0	8.0	43.0	8.5
2R	73.2	7.6	52.6	6.0
3R	75.8	6.6	56.9	7.0

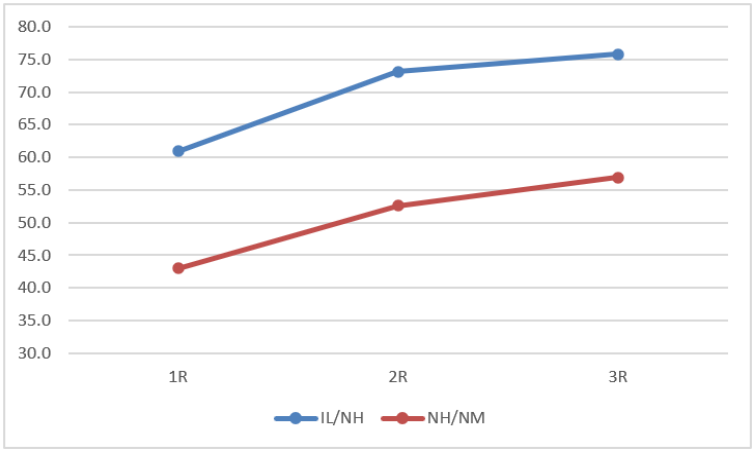

그림 6. 라운드별 분할점수 변화

한편, 수준설정에서 패널별로 산출한 성취수준 분할점수들의 편차를 보여주는 분할점수 표준편차를 살펴보면, 8명이 각각 산출한 IL과 NH 수준을 구분하는 분할점수들의 표준편차가 1라운드에서는 8.0이었으나 3라운드에서는 6.6으로 낮아졌으며, NH와 NM 수준을 구분하는 분할점수의 표준편차는 1라운드에서 8.5였으나 3라운드에서 7.0으로 낮아져, 패널 간 의견의 편차가 줄어들었음을 확인할 수 있었다. 패널들이 개별적으로 추정한 분할점수들의 표준편차 변화를 구체적으로 살펴보면 IL/NH 분할점수의 표준편차는 라운드를 거듭할수록 일관되게 작아지는 경향을 보였으나 NH/NM 분할점수의 표준편차는 2라운드에서 줄어들었다가 3라운드에서 다시 커지는 경향을 보였다. 비록 1라운드에 비해 여전히 표준편차가 줄어들기는 하였으나, 이와 같이 패널 간의 의견차(편차)가 다시 커지는 원인에 대한 면밀한 분석이 이루어질 필요가 있다.

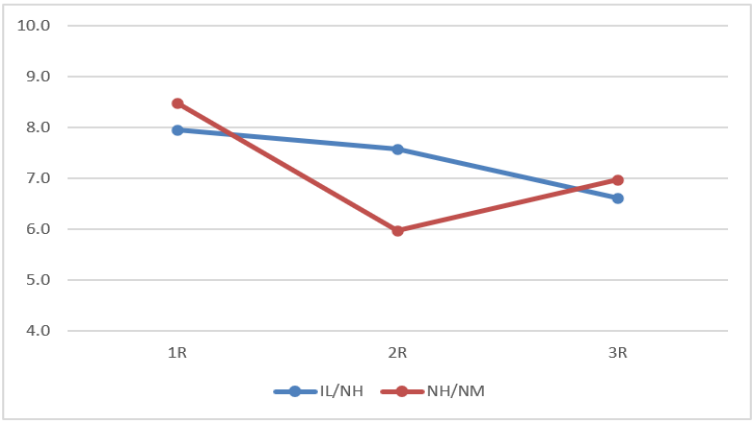

그림 7. 라운드별 표준편차 변화

본 연구에서는 각 라운드 종료 후 문항별로 가장 높은 정답률을 추정한 패널과 가장 낮은 정답률을 추정한 패널의 의견을 들어보고 개별 문항의 특성을 다시 한번 논의한 후, 패널들이 독립적

으로 본인이 추정한 정답률을 다시 한번 검토할 수 있는 시간을 가졌다. 표 4를 살펴보면 패널 D의 경우 1라운드에서 NH/NM의 분할점수를 가장 낮게 추정하였으며(23점) 두 번째로 낮게 추정한 패널 C의 분할점수(38점)과 비교하여 무려 15점의 차이를 보였다. 이후, 1라운드 결과에 대한 검토와 추정한 문항 정답률에 대한 논의를 바탕으로 2라운드를 진행한 결과 패널 D가 추정한 분할점수는 1라운드에 비해 21점(2R-1R)이나 높아져 패널들이 추정한 분할점수의 표준편차를 줄이는 데 기여한 것으로 해석된다. 그러나 여전히 2라운드에서도 패널 D가 추정한 분할점수는 가장 낮은 45점이었다. 이미 상술한 바와 같이 패널들은 라운드를 거듭할수록 NH/NM의 분할점수를 높이는 방향으로 의사결정을 하고 있었으며, 이러한 상황에서 패널들이 2라운드와 3라운드에서 각각 추정한 분할점수의 차이(3R-2R)을 표 4에서 살펴보면, 1라운드부터 안정적으로 분할점수를 산출하고 있는 패널 A를 제외하면 패널 D는 분할점수를 가장 소폭으로 상승시켰음을 알 수 있다. 이미 최종 분할점수에 가장 근접하게 분할점수를 산출하고 있는 패널 A를 제외하면, 2라운드보다 3라운드에서 패널별로 4점에서 8점 정도 분할점수를 높였으나, 패널 D는 2점만 올렸음을 확인할 수 있다. 따라서 2라운드에서 가장 낮게 분할점수를 산출한 패널 D가 최소한 다른 패널과 동일한 편차로 점수를 상승시켜야 했음에도 불구하고 그렇지 못한 것이 전체 표준편차를 증가시키는 결과를 초래한 것으로 해석된다. 최종 분할점수를 평균이 아닌 중앙값으로 결정하기 때문에 패널 D가 추정한 분할점수가 최종 분할점수에 미치는 영향은 없다. 그러나, 향후 안정적인 수준설정을 위해서는 이렇듯 패널별로 산출하는 분할점수의 변화 패턴 등을 면밀해 분석해서 그 결과를 패널들에게 환류하는 과정이 수준설정에 반드시 수반될 필요가 있다.

표 4. 패널별/라운드별 NH/NM 분할점수 변화

패널	1R	2R	3R	2R-1R	3R-2R
A	55	55	56	0	1
B	44	46	51	3	4
C	38	47	53	9	7
D	23	45	47	21	2
E	45	63	68	17	5
F	48	59	67	11	8
G	40	52	60	11	8
H	42	54	58	11	4

수준설정은 문항에 대한 MCP의 정답률을 추정하는 방식으로 분할점수를 산출하지만, 패널들의 정무적 판단도 개입되는 과정이다. 표 3에서 보여주듯이 분할점수가 일관성있게 높아지는 현상을 확인할 수 있다. 이는 각 라운드 종료 후 산출된 분할점수를 적용하여 성취수준별 인원 비율을 계산하고 이러한 성취수준별 비율의 적정성에 대해 패널들이 토론하는 과정에서 영어 시험에 응시한 피험자 집단의 특성과 수준을 고려하여 그림 8과 같이 가장 높은 IL 수준의 비율을 줄이고, 가장 낮은 수준인 NM 순의 비율을 늘리는 방향으로 패널들이 의사결정을 하였음을 알 수 있다.

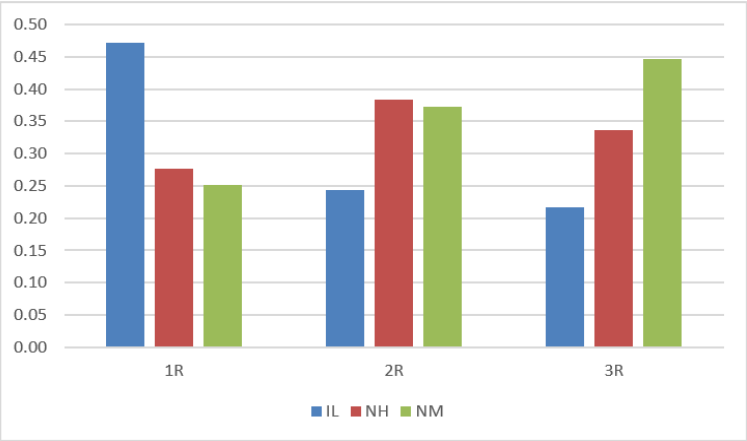

그림 8. 성취수준별 인원 비율

이와 같이 앵고프 방법을 이용하여 3라운드에 걸친 수준설정을 통해 최종적으로 IL, NH, NM의 3가지 성취수준을 구분하는 분할점수를 도출하였으며, 각 수준을 분할하는 점수는 75.8(IL/NH)과 56.9(NH/NM)로 결정되었다. 최종 분할점수를 적용하여 본 영어 시험에 응시한 학생들의 성취수준을 구분한 결과 IL 수준은 77명, NH 수준은 119명, NM 수준은 158명으로 각각 21.8%, 33.6%, 44.6%로 나타났다.

표 4. 수준설정 결과

성취수준	IL	NH	NM	계
분할점수(원점수)	75.8	56.9		
수준인원	77	119	158	354
수준비율	21.8%	33.6%	44.6%	100%

44.6%에 해당하는 NM 수준의 학생들은 영어를 통한 정소 습득이 제한적인 학생들로서 대학 교양영어를 통한 지원이 필요한 학생들이다. 현재까지 A대학에서는 편의상 3개의 분반을 운영하면서 인위적으로 1/3의 학생을 개별 분반에 할당하여 수업을 진행하고 있었다. 따라서 최소 10%이상의 학생들이 적절한 교육적 지원을 받지 못하고 있었음을 알 수 있다. 이와 같이 수준설정을 통해 학생들의 영어 능숙도 수준을 보다 정확히 판정하고 학교 차원에서 학생들의 능력 수준에 맞는 수업운영 방안을 도출할 수 있는 기초자료를 도출할 수 있었다.

5. 요약 및 결론

본 연구는 대학에서 준거참조평가로 시행되는 영어 진단평가를 위한 분할점수 산출 방안을 적용하고, 그 결과를 토대로 대학의 수준별 교양영어 교육의 내실화를 위한 수준설정 방법의 적용 방안을 탐색하였다. 이를 위해 A대학의 교양영어 분반을 위해 사용되던 영어 진단평가 문항을 기준으로 앵고프 수준설정을 실시하여 학생들의 영어 능숙도 수준을 구분하기 위한 분할점수를 산출해 보았다. 수준설정에는 A대학에서 교양영어를 가르치는 내용전문가 8명이 패널로 참여하였다. 본 연구에서는 수준설정을 통해 도출된 분할점수를 A대학의 영어 진단평가에 참여한 354명의 학생들의 진단평가 채점 자료에 적용하여 학생들의 영어 능숙도 수준을 점검하고, 분할점수 산출 과정의 타당성을 검증하는 한편 향후 진단평가에 수준설정을 통한 분할점수 산출이 안정적으로 이루어지도록 하기 위한 시사점을 도출하였다.

본 연구에서 분할점수 산출을 위한 문항 정답률 추정은 3차례 반복 수행하였으며 이를 통해 패널 간의 의견 차이를 좁히는 과정을 거쳤다. 3차례에 걸친 앵고프 수준설정 결과 패널들은 가장 낮은 수준인 NM수준의 비율을 높이고 가장 높은 수준인 IL 수준의 비율을 줄이는 방향으로 분할점수를 산출하는 경향을 보였으며, 라운드를 거듭할수록 패널 간의 의견 편차가 좁아지는 경향을 확인하였다. 구체적으로 살펴보면 각 패널들이 추정한 IL/NH 분할점수의 표준편차는 라운드를 거듭할수록 일관되게 작아지는 경향을 보였다. 이와 비교하여 NH/NM 분할점수의 표준편차는 2라운드에서 줄어들었다가 3라운드에서 다시 커지는 경향을 보였으나 1라운드의 표준편차보다는 여전히 작아 1라운드에 비해 패널 간의 의견 편차가 줄어든 것으로 나타났다. 또한 수준별 최소능력자인 MCP의 정답률 추정치와 이를 종합해 산출된 분할점수들을 분석한 결과 가장 높게 산출된 NH/NM의 분할점수가 대부분의 패널들이 산출한 IL/NH의 분할점수보다 낮아 수준별 MCP에 대해 PLD에서 상정한 이론적인 능력 수준이 실증적으로 구분될 수 있는 능력 수준이며, 동시에 패널별로 MCP에 대한 공감대가 어느 정도 일관성있게 형성되었음을 알 수 있었다.

최종적으로 산출된 분할점수를 적용하여 A대학 진단평가에 응시한 354명의 학생들의 영어성적을 기준으로 영어 능력 수준을 구분한 결과 가장 낮은 수준인 NM 수준으로 판별되는 학생들이 약 44.6%이었으며, 중간 수준인 NH 수준으로 판별되는 학생이 33.6% 그리고 가장 높은 수준인 IL 수준으로 판별되는 학생이 약 21.8%로 나타났다. 현재 A대학에서는 교양영어 수업 운영의 편의상 3개 수준별로 분반을 운영하고 있으며, 23년 2학기의 경우 '상'수준에 6개 분반, '중'수준에 11개 분반, '하'수준 8개 분반으로 실제 성취수준별 비율과 학교에서 운영하는 수준별 분반의 인원 비율에 차이가 있음을 확인할 수 있었다. 이에 따라 A대학에서는 수준에 따라서는 약 10%의 학생들이 본인의 영어 능력 수준과 맞지 않는 분반에서 수업을 받고 있음을 알 수 있었다.

이와 같은 결과는 비단 A대학의 경우에만 해당되는 사례로 볼 수는 없으며, 이른바 고정분할점수를 통해 학생들의 성취수준을 분할하는 대부분의 대학에서는 얼마든지 학생들의 능력 수준과 분반의 수준이 맞지 않는 현상이 발생할 수 있는 가능성이 있다. 따라서 수준별 수업을 지향하는 대학의 경우에 효과적으로 학생들을 지원하고 교양영어 수업을 통해서 학생들의 영어 능력을 증진시키기 위해서는 수준설정의 절차를 통해 산출된 분할점수를 활용하여 학생들의

영어 능력 수준을 구분하고, 이 결과를 바탕으로 수준별 분반을 구성할 필요가 있다.

A대학의 영어 진단평가 수준설정 사례를 분석한 결과, 라운드가 거듭됨에 따라 분할점수에 대한 패널 간의 의견이 좁혀질 것으로 기대하였음에도 불구하고, NH/NM 분할점수의 표준편차 변화와 같이 3라운드의 표준편차가 2라운드보다 커지는 현상을 확인할 수 있었다. 이는 수준설정 이후 진행되는 패널들의 논의 과정에서 개별 패널의 분할점수 산출 경향성을 분석해서 적절한 피드백을 주지 못할 경우, 특정 패널이 산출한 분할점수로 인해 전체 분할점수의 표준편차가 증가될 수 있음을 보여준다. 이는 최종적으로 산출되는 분할점수의 안정성과 신뢰성을 훼손할 수 있는 요인으로 작용할 수 있으므로, 향후 수준설정에서 주의 깊게 살펴보아야 할 부분이라 할 수 있다.

수준설정은 개별 패널들이 PLD와 합의된 MCP를 기준으로 문항을 분석하여 분할점수를 추정하지만, 이 과정에서 정책적 또는 정무적 판단의 개입을 허용하고 있다. 비로 A대학의 사례에서는 이러한 정책적 판단을 배제하였지만, 패널들이 A대학에서 실제로 교양영어 수업을 진행하고 있는 내용전문가들이라는 점에서 이론적으로 정립된 PLD와 MCP 등 이외에도 A대학 학생들의 영어 능력 수준을 고려하여 분할점수를 높이는 방향으로 의사결정 하였음을 알 수 있었다. 그러나 이러한 의사 결정은 자칫 패널들의 전문성과 PLD와 MCP에 기반한 수준설정을 무의미하게 만들 수 있으므로, 향후 수준설정에서는 이러한 정책적 판단의 범위를 명확히 하고 이에 대한 가이드라인을 만들어 수준설정을 진행하는 진행자가 이 가이드라인을 충분히 숙지한 수준설정에 임하도록 할 필요가 있다.

본 연구에서는 A대학의 영어 진단평가의 수준설정 사례 분석을 통해 준거참조평가에서 수준설정의 절차와 고려사항 및 시사점을 살펴보았다. 본 연구에서 사용한 수준설정 방법은 선다형 문항에 최적화된 앵고프 방법을 사용하였으나, 영어와 같은 언어 능력이 선다형 문항으로 평가하기에 용이한 듣기와 읽기 시험만 있지 않고 서답형 문항이 적절한 말하기와 쓰기 시험도 있다는 점을 고려하면, 후속 연구에서는 서답형 문항으로 구성되는 시험에서의 수준설정 절차 및 적용 사례를 분석하여 기존의 선다형 문항 중심의 수준설정과 비교하고 시사점을 도출하고 대학 현장에서 적용 방안을 마련할 필요가 있다.

참고문헌

김성혜, 임자연(Kim, S. H. and J. Lim). 2013. 대학교양영어 프로그램의 운영현황(The current state of college English education in Korea). ≪현대영어교육≫(*Mordern English Education*) 14(2), 263-290.

김성해, 최태환, 사재진, 홍서경 (Kim. S., T. Choi, J. Sa and S. Hong). 2021. 대학영어 단계별 교육과정에 대한 학습자 인식 및 만족도 연구: C대학 사례를 중심으로 (A study of learner perceptions and satisfactions of level-differentiated English education in college: Focused on the case of C university). ≪영어영문학≫(*The Mirae Journal of English Language and Literature*) 26(2), 155-176.

김성숙, 박은아, 서민희(Kim, S., E. Park and M. Seo). 2014. 고등학교 성취평가에서의

고정분할점수와 수준설정방법 적용 결과의 비교 분석(Adoptability of standard-setting methods to the high school level in the achievement standard assessment). ≪교육평가연구≫(*Journal of Eucational Evaluation*) 27(1), 1-22.

김우형(Kim, W. H.). 2015. 수준별 교양영어교육을 위한 영어능력 평가진단 도구개발 연구(A study on the development of an English ability assessment tool for level-differentiated English education). ≪영어영문학연구≫(*Studies in English Language & Literature*) 41(1), 203-226.

곽예린, 김경리(Gwak, Y and K. Kim). 2022. 중학교 수리적성 영역 검사에 대한 Modified-Angoff와 Bookmark 준거설정 방법 비교(A comparison of the modified-Angoff method and Bookmark method for standard setting on formative assessment). ≪교육과학연구≫(*Journal of Educational Studies*) 53(4), 59-82.

박연복, 이규민, 강상진(Park, Y., G. Lee and S. Kang). 2011. 군집분석을 이용한 수준설정 방법과 타당성 연구(A study of standard setting method using the cluster analysis and validity). ≪교육평가연구≫(*Journal of Educational Evaluation*) 24(3), 645-664.

박인용, 송미영, 김성숙(Park, I., M. Song and S. Kim). 2014. 수준설정의 타당성 검증을 위한 평정자 기대 정답률의 예측오차 분석(Accuracy of estimating expected probability for MAP in standard setting). ≪교육과정평가연구≫(*The Journal of Curriculum and Evaluation*) 17(3), 223-245.

서정아(Seo, S). 2018. 대학 교양영어 수준별 수업에 대한 학생들의 인식에 관한 연구(A study of students' perceptions towards level-differentiated general English classes in the university). ≪교육문화연구≫(*Journal of Education & Culture*) 24(5), 355-375.

서혜진(Seo, H.). 2020. 수준별 대학 교양영어에 대한 학습자 만족도 및 인식연구(A study of college students' satisfaction and opinions on level-differentiated instruction of English in general education). ≪교사교육연구≫(*Teacher Education Research*) 59(4), 601-614.

성태제, 시기자, 이경희, 박산하, 권승아(Seong, T., K. Si, K. Lee, S. Park and S. Kwon). 2016. 중학생용 CACV 종합진로직업적성검사 검사 매뉴얼(*Middle School CACV comprehensive career aptitude test examination manual*). 서울: 인싸이트(Seoul: Inpsyt).

오은주(Oh, E.). 2022. 세계시민성 및 글로벌역량 함양을 위한 교양영어 수업운영 사례연구(Integrating global citizenship and global competence into a general English course: A case study). ≪국제이해교육연구≫(*Journal of Education for International Understanding*) 17(1), 93-156.

이영주, 박장희, 정준원, 김수정, 김윤덕(Lee, Y., J. Park, J. Jung, S. Kim and Y. Kim). 2017. 의과대학 수업수준에서의 준거설정 사례 연구(Study on standard setting methods applied to medical school class). ≪교육문화연구≫(*Journal of Education & Culture*) 23(2), 189-209.

이선경, 지은림(Lee, S and E. Ji). 2017. Rasch 방법과 Extended-Angoff 방법을 적용한 인성검사 준거설정 비교 연구(Comparing Rasch method and extended-Angoff method in standard setting for character Test). ≪교육평가연구≫(*Journal of Educational Evaluation*) 30(4),

761-789.

임의진, 이용원, 전희성(Lim, E., Y. Lee and H. Jeon). 2019. 유럽 언어 공통기준 연계를 통한 영어 독해 시험의 수준설정: 수정된 앙고프 방법과 북마크 방법 비교(Standard setting to relate an English reading comprehension test to the CEFR: A comparison of modified Angoff and Bookmark methods). ≪교육평가연구≫(*Journal of Educational Evaluation*) 32(3), 523-548.

정양수(Jung, Y.). 2014. 교양영어 수준별 수업의 운영 효과 및 개선 방안에 관한 연구(A study on the implementation of level-specific general English and its effects). ≪언어학연구≫(*Journal of Linguistic Studies)* 19(3), 127-148.

Educational Testing Service. 2021. *Mapping the Redesigned TOEIC Bridge Test Scores to Proficiency Levels of the Common European Framework of Reference for Languages*. Princeton, NJ: Educational Testing Service.

Zieky, M. J., M. Perie and S. A. Livingston. 2008. *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Examples in: English, Korean
Applicable Languages: English, Korean
Applicable Level: Tertiary

