*영어어문교육 16*권 **4**호 **2010**년 겨울

국가영어능력평가시험(NEAT)의 검사지 구성의 원칙과 절차: 문항 유형 확정 모델

김 용 명
(한국교육과정평가원 KICE)

Kim, Yong-Myeong. (2010). A blueprint for designing and developing the listening and the reading test of National English Ability Test (NEAT): Item-types decision-making model. *English Language & Literature Teaching, 16*(4), 153-184.

On the bases of the 5 principles and the 4 criteria for designing and developing of the listening and the reading test of National English Ability Test (NEAT), this study presents Item-Types Decision-Making Model as a blueprint for designing and constructing the two tests. It sets up the criteria for validating item types, designs a *modular* type of test specifications, constructs an item-types bank, and specifies a *complementary* type of test specifications of the two tests. To gather all these threads up, it constructs Item-Types Decision-Making Model which consists of such components as the item-type pool, the validity criteria and the procedures of testing item types, the item-types bank, the modular and the complementary type test specification. Thus, it shows how the Model works in developing and constructing the two level-differentiated listening and reading tests (the 2nd and the 3rd rank) of NEAT. Finally, it discusses some implications and applications of the Model to the two level-differentiated tests (the A and the B type) of 2014 CSAT (College Scholastic Ability Test) systems, National Assessment of Educational Achievement (NAEA), and classroom testing. In conclusion, Item-Types Decision-Making Model functions as a testing template in an item development system and as a matrix in an item-types bank system.

[National English Ability Test (NEAT)/item type/modular · complementary type test specification, 국가영어능력평가시험/문항 유형/문항 유형 은행/모듈형 · 상보형 평가목표 이원분류표]

Ⅰ. 서론

2010 년대 한국 영어 평가에는 세 가지 주요 변화가 예고되어 있다. 이 변화에는 대학수학능력시험(이하, 수능) 외국어 영역의 듣기 문항 확대 방안,

2014 수능 체제 개편(안), 국가영어능력평가시험(National English Ability Tests, 이하 NEAT)의 수능 영어 시험 대체(안) 등이 있다.

수능 외국어(영어) 영역의 듣기 문항 확대 방안은 개정 영어과 교육과정(교육인적자원부, 2007a)을 수능 외국어(영어) 영역에 반영하고, 현행 외국어(영어)영역 체제에서 듣기와 읽기 영역 간의 불균형(34%:66%)을 해소하기 위해, 듣기 문항의 수를 현행 17(34%)문항에서 25(50%)문항으로 확대하는 것을 주요 내용으로 한다(대통령 업무 보고, 2009.12.23; 김용명, 고현숙, 김진석, 이완기, 2010). 또한 2014 수능 체제 개편(안)은 수준별 수능과 복수 시행을 주요 골자로 한다(중장기 대입선진화 연구회, 2010). 2014 수능 체제에서는 현행 수리 영역(수학)과 같이, 영어도 두 수준, 즉, A 형과 B 형의 시험을 제공하며, A 형은 현행 수능보다 낮은 수준으로 출제 범위는 줄이고, 보다 쉽게 출제하여 수험생의 부담을 최소화하며, B 형은 현행 수능 수준과 동일한 출제 범위와 난이도를 유지한다. 이에 따라 수험생은 자신의 수준과 진로에 따라 A 형 또는 B 형을 선택하여 시험을 볼 수 있고, 또한 복수 시행 체제에 따라 수험생은 연 2 회 응시할 수 있다. 2014 수능 체제 개편(안)이 2010 년 하반기에 확정되면, 듣기 문항 확대 방안과 맞물려 수준별(A·B 형) 시험의 출제 범위, 시험 체제, 평가 내용, 문항 유형 등을 재설정할 필요가 있다.

NEAT 의 수능 대체 여부는 2010 년대 한국 영어 평가 변화의 최종 귀착점이 될 것이다. 교육부 발표(교육인적자원부, 2006, 2007b, 2008b)에 따르면, NEAT 는 듣기, 읽기, 말하기, 쓰기 4 개의 평가 영역으로 구성되며, 각 영역은 수준별, 1 급, 2 급, 3 급으로 구성된다. 1 급은 성인용으로 대학 2-3 학년 수준으로 대학 졸업이나 취업 시에 토익, 토플 등을 대체하는 용도로, 2·3 급은 학생용으로 대입 전형 자료로 활용될 것이다. 또한 NEAT 는 IBT(Internet-Based Tests, 이하 IBT) 방식의 복수 시행 체제로 운영될 것이다. 최근 교육과학기술부(2010) 발표에 따르면, NEAT 의 수능 외국어(영어) 시험 대체 여부는 여론 수렴 과정을 거쳐 2012 년에 최종 결정하여, 2016 학년도부터는 수능 영어 시험을 대체하는 방안이 추진되고 있다. 이에 따라 NEAT 의 수준별(2·3 급), 기능별(듣기·읽기·말하기·쓰기) 출제 및 시행 체제, 평가 내용, 평가 요소 및 문항 유형 등을 확정할 필요성이 있다.

2010 년대 영어 평가의 세 가지 변화가 교육적으로 의미 있고 바람직한 방향으로 변화가 될 것인지, 이에 따라 세 변화의 최종 정점에 있는 NEAT 가 진정한 의미에서 수능의 창조적 계승자가 될 것인지 아니면 수능의 단순한 대체자에 머무를 것인지는 결국 어떤 원칙에 따라 어떤 문항 유형을 어떻게 조합하여 NEAT 에 타당한 검사지를 구성할 수 있느냐에 달려있다. 그러나 현재까지의 선행 연구를 살펴보면, 교실 평가는 물론 수능, NEAT 등과 같은 대단위 고부담 시험과 관련된 문항 유형 개발 원리 및 검사지 구성의 원칙에

관한 연구는 찾아 볼 수 없었다[1]. 이와 같은 연구 공백을 메우고, 또한 NEAT 의 구성에 시사점을 제공하기 위해, 김용명(2010)은 NEAT 설계의 밑그림으로서 '문항 유형 결정 원리'와 '검사지 구성 원칙'을 제안했다. 문항 유형 결정 원리는 연계성(connection), 실제성(authenticity), 상호작용성(interactiveness), IBT 양립성(Compatibility), 환류 효과성(washback effectiveness) 등으로 구성되며, 개별 문항의 개발 및 선별에 관계한다. 한편 검사지 구성 원칙은 상보성(complementarity), 통합성(integration), 주축성(pivotality), 위계성(hierarchicality) 등으로 구성되며, 문항 유형 결정 원리에 따라 개발 또는 선별된 문항들 간의 상호 관계에 작동한다.

본 연구는 이에 대한 후속 연구로서 NEAT 설계의 밑그림으로 제안된 문항 유형 결정 원리와 검사지 구성의 원칙을 토대로 NEAT(듣기·읽기)의 실행 설계도(blueprint)를 제안하고자 한다. 이 같은 연구 목적을 구체화하기 위하여, 본 연구에서는 NEAT 의 설계도의 주요 구성 요소를 다음과 같이 연구 주제로 설정한다.

(a) NEAT 의 시행 및 시험 체제의 구성
(b) 문항 유형 풀(Item-Types Pool)의 구성
(c) 문항 유형 타당성 평가 기준과 점수화 절차의 설정
(d) 문항 유형 은행(Item-Types Bank)의 구성
(e) 모듈형(modular)과 상보형(complementary) 평가목표 이원분류표의 구성

이 같은 연구 주제에 대한 가능한 안을 탐색하기 위해, 2 장에서는 현재 개발 중에 있는 NEAT 시험의 개요를 설명하고, 또한 NEAT 설계도의 이론적 틀로서 문항 유형 결정 원리와 검사지 구성 원칙에 관해 설명할 것이다. 이를 토대로 3 장에서는 NEAT 의 시험 체제의 구성에 관해 논의할 것이며, 4 장에서는 NEAT 검사지 구성의 구체적 절차와 방법에 관한 일종의 흐름도(flowchart)로서 '문항 유형 확정 모델(Item-Types Decision-Making Model)'을 제안할 것이며, 마지막 5 장에서는 NEAT 의 설계도로서 제안한 문항 유형 확정 모델이 2014 수능 체제의 수준별 영어(A·B 형), 국가수준학업성취도평가 및 교실 평가에 줄 수 있는 이론적 실제적 시사점에 관해 논의할 것이다.

[1] 수능 외국어(영어)영역의 읽기의 난이도 분석(성윤미, 2003; 장경숙, 2004), 읽기의 지문 분석(신명신, 1999), 지문 친숙도 분석(이경숙, 1999), 어휘 분석(김낙복, 2008) 등과 같은 단편적인 연구만 찾을 수 있었다.

Ⅱ. 국가영어능력평가시험(NEAT)의 검사지 구성의 이론적 틀

1. 국가영어능력평가시험(NEAT)의 개요

2010 년대 한국 영어 교육의 최대의 관심사는 2012 년에 결정될 NEAT 의 수능 외국어(영어)영역 대체 여부일 것이다. 이에 본 절에서는 NEAT 의 취지와 목적, 성격, 시험 및 시행 체제, 평가 체제 등에 관해 간략히 설명하고자 한다.

교육과학기술부(2007b, 2008b, 2010)의 NEAT 추진 경과에 따르면, NEAT 의 목적과 취지는 국가 주도의 영어 평가 시험을 개발하여 국내용 목적의 해외 시험(TOEIC 등)을 대체함으로써 해외 영어 시험에 대한 의존도를 낮추는데 있으며, 또한 학생들의 수준과 장래 진로에 따른 수준별(2·3 급) 영어능력 평가 시험을 개발함으로써 학교 영어 교육의 방향에 대한 바람직한 기준을 제시하는데 있다. 또한 NEAT(2·3 급)는 개정 영어과 교육과정에 따른 듣기, 읽기, 말하기, 쓰기의 세부 성취기준 달성 정도를 측정함과 동시에 학생의 일반 영어 능력을 평가한다. 이런 점에서 NEAT 는 성취도 시험의 성격과 숙달도 시험의 성격을 모두 갖는다(한국교육과정평가원, 2010b, p. 7).

표 1

NEAT 2.3급 듣기 및 듣기 시험의 평가 체제

		듣기 2급	듣기 3급	읽기 2급	읽기 3급
듣기 대화문	단어 수	비공개	비공개	-	-
	Turn-taking 수	비공개	비공개	-	-
듣기 담화문	1지문 1문항 단어수	비공개	비공개	-	-
	1지문 2문항 단어수	비공개	비공개	-	-
읽기 지문	단문 독해 단어수	-	-	비공개	비공개
	장문 독해 단어수	-	-	비공개	비공개
선택지 (공통)		4지선다형 (일부 3지)	4지선다형 (일부 3지)	4지 선다형	4지 선다형
속도 (공통)		비공개	비공개	-	-
총 문항 수		35	35	35	35
시험 시간		35분	35분	60분	60분

국가영어능력평가시험(2·3 급) 개발 및 운영 방안을 주제로 한 공청회 자료(한국교육과정평가원, 2010b)에 따르면, NEAT 2 급과 3 급의 듣기 및 읽기의 평가 체제는 표 1 과 같다. 듣기 2 급과 3 급의 총 문항 수는 35 개로

구성되어 있으며, 시험 시간도 35 분씩 동일하다. 5 지 선다형으로 구성된 수능과 달리 NEAT 는 4 지 선다형으로 구성되어있다(관용적 응답 유형의 경우, 3 지 선다형). 한편 읽기 2 급과 3 급의 선택지는 듣기와 마찬가지로 4 지 선다형으로 구성되며, 총 문항 수는 35 개이며, 시험 시간도 60 분씩 동일하다.

마지막으로 NEAT 의 출제 체제 및 시행 체제를 살펴보고, 이와 관련된 시사점을 살펴보자. 이미 언급한 바와 같이, NEAT 는 수준별(2·3 급) 평가 제제와 복수 시행 체제를 따를 것이며, IBT 방식으로 시행될 예정이다. 수준별 평가 체제에 따라 학생들은 자신의 영어 능력 수준과 진로에 따라 2 급과 3 급을 선택할 수 있으므로, 학습 부담을 경감시킬 수 있는 장점이 있다. 또한 복수 시행 체제에 따라 학생들은 연간 수회(2-3 회)에 걸쳐 NEAT 에 응시할 수 있음으로써 시험 당일의 '실수' 에 대한 심적 부담을 줄일 수 있을 뿐만 아니라 수시로 응시하여 자신의 성적을 점검할 수 있어 자기 주도적 학습을 할 수 있을 것이다. 또한 NEAT 의 IBT 방식 시행 체제는, Roever(2001), Chapelle 과 Douglas(2006) 등이 언급한 바와 같이, 시간과 장소의 자율성 및 편리성이 있을 뿐만 아니라 IT 기술과 접목된 실제성과 상호작용성이 높은 새로운 유형의 문항의 개발을 촉진시킬 것이며, 이에 따라 교실 수업에서도 IT 기술과 융합된 창의적인 학습 과업 및 학습 자료의 개발이 촉진될 것이다.

그러나 NEAT 가 수준별(2·3 급)로 복수 시행 체제에 따라 IBT 방식으로 실제 시행된다고 가정해보면, 해결해야 할 문제점이 적지 않음이 드러난다. NEAT 의 응시생 수를 60 만(2010 학년도 수능 기준)으로 상정하고, 수험생 당 최소 년 2 회 응시 기회를 준다고 전제하면, 수험생 수는 연간 120 만 명을 넘을 것으로 예측된다. 전국적으로 5 만개의 IBT 시험장을 상정하면, 연간 24 회 이상의 NEAT 가 시행될 것으로 추산된다. 이 같은 시뮬레이션을 통해 드러난 바와 같이, NEAT 의 출제 체제는 현행 수능에서 시행하고 있는 것과 같은 장기 폐쇄형 합숙 출제는 불가능할 것이며, 문제 은행식 또는 문항 공모식 출제 체제가 불가피할 것으로 생각된다. 문제 은행 출제 체제를 따를 경우, 매 시행되는 시험 마다 적시에 적소의 문항을 축출할 수 있고, 또 재고 문항과 부족 문항을 상시적으로 파악하여 적시에 부족 문항 개발하여 보충할 수 있는 포괄적이면서도 정교한 문제은행 시스템을 구축해야 할 것이다. 본 연구에서는 단순히 문항을 저장하고 축출하는 기존의 문제 은행 체제 대신에 '모듈형' 평가 목표 이원분류표에 따른 매트릭스 방식의 '문항 유형 은행(item-types bank)' 체제를 제안할 것이다(다음 4 장 참조). 또한 연간 24 회나 시행되는 복수 시행 체제를 따를 경우, 매 시험 간 항상성과 동등성을 유지하여 시험의 공정성을 확보하는 것이 무엇보다 중요하므로 매 시행되는 시험에서 나온 점수를 호환적으로 사용하기 위한 동등화(equating) 과정이 필수적이다. 그러나 이에 앞서 지난 20 년간 시행되어온 수능 문항

분석 자료와 NEAT 예비 시행 분석 자료를 토대로 각 문항 유형들이 갖는 특성을 심층 분석하여 시험의 항상성과 동등성에 기여할 수 있는 문항으로 NEAT 의 검사지를 구성하는 것이 바람직할 것이다. 이를 위해 본 연구는 각 개별 문항 유형을 그 특성에 따라 분류한 '문항 유형 특성 체계' 에 따라 NEAT 의 검사지를 구성할 것을 제안할 것이다(다음 4 장 참조).

마지막으로 NEAT 는 전국적으로 시행되는 고부담 시험이며, 더구나 최초로 IBT 로 시행된다는 점에서 시험의 형평성과 민주성의 확보가 그 어느 시험보다 중요하다. 따라서 IBT 양립성 원리에 입각해 컴퓨터 및 인터넷 친숙도에 따라 지역·계층 간 시험 수행(점수)의 차이가 생기지 않도록 PBT(Paper-Based Tests, 이하 PBT)와 IBT 간의 동질성 연구의 결과를 반영하여 검사지를 구성해야 할 것이다. 예를 들어, 동질성 분석 결과, 지역·계층 간 시험 수행(점수)의 차이가 상대적으로 작은 문항 유형은 검사지에 그대로 포함될 수 있지만, 시험 수행의 차이가 상대적으로 큰 문항은 그 요인을 면밀하게 분석하여 검사지에 포함할 것인지를 결정해야 할 것이다. 가령, 두 시험 간의 수행의 차이가 컴퓨터 친숙도 차이에서 기인된 것인지 아니면 PBT 에서 IBT 로 구현되면서 문항의 실제성과 상호작용성의 차이에서 기인된 것인지 식별해야 할 것이다. 전자의 경우, 수험생의 언어 능력 이외의 요소가 시험 수행 결과에 반영된 것이므로 검사지에서 배제되어야 한다. 그러나 후자의 경우, 구현성 및 작동의 편의성을 재고하거나 NEAT 튜토리얼(tutorial)을 제공함으로써 수행의 차이를 완화할 수 있다면 검사지에 포함될 수 있을 것이다. 특히 듣기나 읽기 입력의 속도는 지역·계층 간 시험 수행의 차이에 민감하므로 신중하게 입력의 속도를 결정해야 할 것이다(다음 절 참조)

2. 문항 유형 개발과 검사지 구성의 절차와 방법에 관한 이론적 틀[2]

김용명(2010)은 NEAT 구성에 관한 밑그림을 제시하기 위해, "어떤 원리에 따라 NEAT 에 적합한 문항 유형을 개발하고, 또 어떤 원칙에 따라 어떤 문항 유형으로 NEAT 에 타당한 검사지를 구성할 것인가?" 라는 연구 질문을 제기하고, 이에 대한 가능한 답으로서 '문항 유형 결정 원리' 와 '검사지 구성 원칙' 을 제안하였다.

문항 유형 결정 원리는 Bachman 과 Palmer(1996)의 시험 유용성(test usefulness) 모델을 토대로 구성한 것으로 연계성(connection), 실제성(authenticity), 상호작용성(interactiveness), IBT 양립성(Compatibility),

[2] 본 절에서 NEAT 구성의 이론적 틀로서 제시한 문항 유형 결정 원리와 검사지 구성 원칙은 본 연구의 선행 연구인 김용명(2010)의 연구 내용을 요약한 것이다.

환류 효과성(washback effectiveness) 등으로 구성된다(김용명(2010, pp. 377-385).

연계성은 NEAT 의 평가 목표, 평가 내용, 수준 및 평가 요소는 개정 영어과 교육과정(교육과학기술부, 2008a)의 교육 목표, 교육 내용, 성취 기준과 연계되어야 한다는 원리를 말한다. 따라서 연계성 원리는 NEAT 의 수준별(2·3 급), 기능별(듣기·읽기) 시험의 출제 체제 구성에 논리적 타당성을 제공하며, 특히 각 수준별 시험의 출제 범위, 평가 내용, 평가 기준 설정에 논리적 근거를 제시해 준다. 실제성은 시험 과업(test task)의 특성과 실제 언어 사용 상황에서의 TLU(Target Language Use, 이하 TLU) 과업의 특성은 상호 일치해야 한다는 원리로 문항 유형 개발 및 선별의 범위와 한계를 설정하는 역할을 한다. 따라서 실제성 원리는 NEAT 의 평가 목표 이원분류표 상의 내용 영역 및 행동 영역 체계의 주요 구성소를 설정하고, 이의 타당성을 검증하는 수단이 된다. 상호작용성은 시험 과업과 학습자의 언어능력이 상호작용하는 정도를 말하는데, 상호작용성이 높은 문항일수록 언어능력과 관계하는 정도가 높으므로 상호작용성 원리는 문항의 질적 통제 역할을 한다. IBT 양립성은 IBT 환경과 수험자 간에는 친화성이 있어야 하며, IBT 환경과 IBT 문항 간에는 IT 기술의 구현성이 있어야 하며, 수험자와 IBT 문항 간에는 언어능력과 상호작용성이 있어야 한다는 원리를 말한다. 따라서 IBT 양립성 원리는 IBT 문항의 기술적 통제 역할을 한다. 마지막으로 환류 효과성은 시험의 시행 결과는 교육과정, 교수·학습 활동, 평가에 긍정적 영향을 주어야 한다는 원리로 교육과정, 교수·학습, 교육평가 간의 괴리를 줄이고, 일체화에 기여한다.

검사지 구성 원칙은 문항 유형 결정 원리와 지난 6 년간 수능 외국어(영어) 영역의 문항 분석 자료(한국교육과정평가원, 2005b, 2006, 2007, 2008, 2009, 2010a)의 심층 분석을 통해 설정한 것으로 상보성(complementarity), 통합성(integration), 주축성(pivotality), 위계성(hierarchicality) 등으로 구성된다(김용명, 2010, pp. 386-392).

상보성은 평가의 모든 요소, 모든 영역, 모든 내용과 그 하위 요소들은 서로 상보적 관계(complementary relationships)에 있어야 한다는 원칙을 말한다. 환언하면, 상보성 원칙은 검사지를 구성함에 있어서 어떤 문항 유형이 필요 불가결한 문형이며, 또 몇 문항이 필요 최소한의 문항인가를 결정하는 역할을 한다. 따라서 상보성 원칙은 필요 불가결한 문항을 필요 최소한으로 검사지를 구성해야 한다는 검사지의 질과 양을 결정하는 원칙이라고 할 수 있다. 또한 상보성 원칙은 동형 또는 평형 문형에 대해서는 상호 배타성을 가지므로 시험의 동등화와 관련하여 동형 검사형(alternative forms)을 구성하는 실행 장치(executive device) 로서 기능을 한다.

통합성은 언어 능력은 구분가능(divisible)하다는 전제에 따라 각 기능별 시험(듣기, 말하기, 읽기, 쓰기)은 해당 기능의 고유한 능력만을 측정하는 기능 독립형 문항만으로 검사지를 구성해야 한다는 논리와 최근 교수이론(Bachman 과 Palmer, 1996; Brown, 2007; Ellis, 2003)과 영어과 교육과정(교육과학기술부, 2008a) 에서 강조되고 있는 4 기능의 통합 교수원리를 토대로 구성한 개념이다. 따라서 통합성 원칙은 각 기능별 시험은 기능 독립형 문항으로 검사지를 구성하되, 기능 통합형(예, 듣기와 말하기 통합) 또는 기능 연계형 문항(읽기와 쓰기를 연계)도 필요 최소한으로 검사지에 포함되어야 한다는 원칙을 말한다.

수능 외국어(영어) 영역의 문항 분석 자료(한국교육과정평가원, 2005-2010)에 따르면, 문항 유형을 주축 문항(pivot item)과 주변 문항(peripheral item) 으로 구분할 수 있다. 주축 문항은 평균 정답률을 기준으로 정답률의 편차(variation)가 상대적으로 작으며, 문항의 복잡도(어휘적, 언어적, 개념적, 인지적 복잡도)가 높아지면, 이에 따라 난이도도 올라가는 경향을 갖는다(문항의 복잡도와 난이도는 정비례 관계). 이런 경향성에서 주축 문항은 학습자의 언어능력에 지배를 받는다는 것을 추론할 수 있으며, 시험의 항상성 유지와 동등화에 기여할 수 있다는 점을 예상할 수 있다. 반면 주변 문항은 정답률의 편차가 상대적으로 크고, 문항의 복잡도가 높아지면, 일정 수준까지 난이도가 올라가지만, 그 수준 이상에서 더 이상 난이도의 변화가 없는 시험 고원(testing plateau) 현상이 생겨나는 경향을 보인다. 이런 점에서 주변 문항은 언어능력보다는 시험 요령, 학습자의 정의적 특성, 배경 지식 등의 지배를 받는다는 것을 추론할 수 있으며, 시험의 다양성에 기여할 수 있다는 점을 예측할 수 있다. 따라서 주축성은 주축 문항을 시험의 항상성과 동등성을 유지할 만큼 필요 최대한으로, 시험의 다양성을 해치지 않을 만큼 필요 최소한으로 검사지 구성에 포함해야 한다는 원칙을 말한다.

위계성은 Kim(2006, 2007a, 2007b)의 시험가능성(testability)[3]의 개념과 수능 외국어(영어) 영역의 문항 분석 결과를 토대로 구성한 개념으로 검사지를 구성하고 있는 각 문항의 복잡도 또는 난이도는 각 수험자(또는 각 집단)의 수행가능 단계와 일치할 수 있도록 위계화되어야 한다는 원칙을 말한다. 외국어(영어) 영역의 문항 분석에 따르면, 각 문항 유형에 대한 수험자 (학습자)의 문항 반응 곡선은 해당 문항 유형의 특성에 따라 L 형, M 형, H 형으로 정형화할 수 있다[4]. L 형 문항 반응 곡선은 하위 등급(예, 9,

[3] The Testability hypothesis was extrapolated from what Pienemann calls Teachability principle (1985, 1998). Given Teachability principle, then, any test task will be 'performable' to the test taker if it relates to structures or rules of the next or subsequent stage of the IL [interlanguage] learners' current stage (Kim, 2007, p. 51).

[4] 문항 반응 곡선 L 형, M 형, H 형은 각각 low, mean, high 의 첫 글자를 딴 것으로 하위 학습자, 중위 학습자, 상위 학습자를 보다 잘 변별한다는 의미를 담고 있다.

8, 7 등급)에서는 등급이 올라감에 따라 정답률도 올라가지만(등급과 정답률이 정비례 관계에 이지만), 일정 등급(예, 7 등급) 이상을 넘어서면 정답률(예, 80% 대)이 고정되어 시험 고원을 형성하는 문항 유형을 말한다. L 형 반응 곡선의 특성을 보이는 문항 유형에는 *a, b, c* 등의 유형이 있으며[5], 이 유형들은 하위 등급 간의 난이도 및 변별도 조정에 결정적 역할을 할 것이다. 한편 M 형의 문항 반응 곡선은 하위 등급(예, 7 등급 이하)에서 정답률이 정체(예, 20% 대)되어 시험고원을 형성하지만, 일정 등급(예, 7 등급)에 도달하면 등급이 올라감에 따라 정답률이 올라간 후, 다시 일정 등급(예, 3 등급)을 넘어서면 정답률이 정체(예, 80% 대)되어 또 하나의 시험 고원을 형성하는 유형을 말한다. M 형 곡선의 특성을 보이는 문항 유형에는 *o, p, q, r* 등의 유형이 있으며, 이런 문항 유형들은 중위 등급 간의 난이도 및 변별도 조정에 결정적 역할을 할 것이다. 반면, H 형의 문항 반응 곡선은 일정 등급(예, 3 등급)까지는 정답률이 정체(예, 30%대)되어 시험 고원을 형성하지만, 일정 등급(예, 3 등급)을 넘어서면 등급과 정답률이 정비례 관계에 있는 유형을 말한다. H 형의 문항 반응 곡선의 특성을 보이는 문항 유형에는 *x, y, z* 등의 유형이 있으며, 이들 문항 유형은 상위 등급 간의 난이도 및 변별도 조정에 결정적 역할을 할 것이다. 따라서 위계성은 한 검사지에서 L 형, M 형, H 형의 구성 비율을 적절히 조정함으로써 검사지 총체적 난이도와 변별도를 적정 수준으로 통제할 수 있고, 더 나아가 시험의 안정성과 항상성을 유지할 수 있고, 시험의 동등화에도 기여할 수 있다. 예를 들어, L 형에 속하는 문항 유형(a, b, c)의 비율을 늘리면, 검사지의 총체적 난이도는 내려갈 것이며, 그 결과 하위 학습자(수험자)에 대한 변별력이 높아지는 경향을 보일 것이다. 역으로 H 형에 속하는 문항 유형(x, y, z)의 비율을 늘리면, 검사지의 총체적 난이도는 올라갈 것이며, 그 결과 상위 학습자(수험자)에 대한 변별력이 높아지는 경향을 보일 것이다.

이상에서 살펴본 바와 같이, 문항 유형 결정 원리와 검사지 구성의 원칙은 상호 배타적 독립적 관계에 있다기 보다는 상호 보완적 유기적 순환 관계에 있다. 다음 장에서는 NEAT 의 밑그림으로 제안한 문항 유형 결정 원리와 검사지 구성의 원칙이 어떻게 서로 상호 보완적 유기적 순환 관계를 통해 NEAT 의 실행 설계도의 주요 구성소, 즉, 출제 및 시행 체제, 문항 유형 풀, 문항 유형 타당성 평가, 문항 유형 은행 체제, 평가 목표 이원분류표 등을 설계하는지를 구체적으로 논의하고자 한다.

---

[5] 수능 외국어(영어) 영역 시험(2005-2010)의 문항 분석을 통해, L 형, M 형, H 형에 속하는 문항 유형을 식별해낼 수 있었지만, 수능 자료가 보안인 관계로 이를 구체적으로 명시하지 않고, *a, b, c* 등으로 표시하였다.

Ⅲ. 국가영어능력평가시험(NEAT)의 시험 체제 구성

1. 국가영어능력평가시험(NEAT)의 출제 범위 및 기준 설정

앞서 논의한 바와 같이, NATE(2·3 급)은 숙달도 시험의 성격과 성취도 시험의 성격을 동시에 가진다. 따라서 언어능력에 대한 이론적 모델(예, Communicative Language Abilities(Bachman, 1990))에 따라 시험 과업과 시험을 구성하되, 출제 범위와 평가 내용은 영어과 교육과정의 범위 안에 있어야 한다. 연계성 원칙에 따라 NEAT 의 수준별(2·3 급), 기능별(듣기·읽기) 시험의 출제 범위 및 출제 과목이 결정될 수 있을 것이며, 이에 따라 각 시험에서 사용될 수 있는 총 어휘 수, 지문 또는 대화(담화)문 당 적정 단어 수가 결정될 수 있을 것이다.

외국어과 교육과정(교육과학기술부, 2008a)은 국민 공통 기본 교육과정과 선택 교육과정으로 구성되어있다. 선택 교육과정은 영어 I, 영어 II, 실용 영어 회화, 심화 영어 회화, 영어 독해와 작문, 심화 영어 독해와 작문 등의 과목으로 구성되어 있다. NEAT 3 급이 실용적 성격의 영어 능력을 평가하고자 한다면, 이에 대응하는 시험 과목은 공통 교육과정의 영어와 심화 선택과정의 실용 영어 회화로 하고, 어휘 수는 영어를 기준으로 1800 단어 내외로 제한하는 것이 바람직할 것이다. 반면, 2 급이 학문적 성격의 영어 능력 평가에 초점을 둔다면, 이에 대응하는 시험 과목은 심화선택 과정의 영어 I, 영어 II, 영어 독해와 작문으로 하고, 어휘 수는 영어 II 를 기준으로 2800 단어 이내로 한정하는 것이 타당할 것으로 생각한다[6]. 심화 영어 회화와 심화 영어 독해 작문은 AP(Advanced Program) 과정으로 두는 것이 진정한 의미에서 '심화' 교육과정에 부합할 것으로 생각한다. 2010 년대 영어교육의 변화의 귀착점에 있는 NEAT 의 수능 대체 과정이 연착륙될 수 있도록 하기 위해서는 NEAT 의 출제 범위 및 평가 요소는 2014 수능 체제의 수준별 영어(A·B 형)와 연계되어야 할 것이다.

2. 국가영어능력평가시험(NEAT) 시험 체제 구성

NEAT 의 검사지를 구성하기 위해서는 NEAT 의 취지, 성격, 목표, 용도에 부합하는 시험 체제를 구성해야 한다. 앞서 논의한 바와 같이, NEAT 는 수준별(2·3 급)로 복수 시행 체제에 따라 IBT 방식으로 시행될 것이다. 이 같은 점을 고려하여 NEAT 의 시험 체제 구성을 위한 시사점을 얻기 위해 먼저 국·내외 표준화 시험을 선정하여 이를 분석해야 한다. 분석 대상

[6] 이 같이 시험 과목을 설정할 경우, 3 급은 6 차 교육과정에 따라 공통영어를 출제 범위로 했던 2005 학년도 이전 수능 체제와 동일할 것이며, 2 급은 7 차 교육과정에 따라 심화학습 과정까지를 포함하는 현행 2005 수능 체제와 유사할 것이다.

표준화 수행력 시험에는 ETS 에서 개발 시행하고 있는 TOEFL 과 TOEIC, 영국과 호주에서 개발 시행하고 있는 IELTS, 한국과 비슷한 교육적 환경을 지닌 일본의 EIKEN, 중국의 CET, 한국의 TEPS, FLEX, MATE, 동양 3 국의 대학입학시험(수능, 중국고시, 일본입시센터시험) 등이 포함될 수 있다.

이와 같은 국·내외의 주요 표준화 시험의 시험 체제를 Bachamn 과 Palmer(1996)가 제안한 시험 양상(test method facets) 틀에 따라 분석한다. 이 분석 결과를 연계성, 실제성, 상보성, IBT 양립성 원리에 비추어 그 타당성을 검증해 본다. 가령, 연계성과 실제성의 원리에 따라 NEAT 의 평가 내용이 결정될 것이며, 상보성 원칙에 따라 필요 최소한의 문항 수가 결정될 것이다. 또한 이를 토대로 수준별 기능별 각 시험의 적정 시험 시간도 추산될 수 있을 것이다. 이와 같은 실증적, 이론적 분석을 통해, 표 2 에서 보는 바와 같이, NEAT 의 시험 체제의 주요 구성소인 적정 문항 수, 시험 시간, 배점, 선택지 수 등을 결정할 수 있을 것이다.

표 2

NEAT 의 시험 체제 및 문항 구성 요소

시험 체제 및 문형 구성 요소		듣기		읽기	
		2 급	3 급	2 급	3 급
시험체제	총 문항 수				
	배점(차등배점)				
	선택지의 수				
	총 시험 시간				
문항구성요소	문항 제시 방법(시각, 청각 등)				
	문항의 제시 순서				
	문항 당 시간 자율성				
	문항 간 이동 여부				
	입력과 응답 방식				
	입력과 응답 속도				
	답안 수정 여부				

한편 IBT 양립성 원칙에 따라 환경 타당도(context validity) 검증을 통해 문항 구성 요소를 설정할 수 있을 것이다(Weir, 2005)[7]. 먼저 Bachamn 과 Palmer(1996)시험 양상을 분석틀로 삼아 PBT 와 IBT 시험 양상을 상호 비교하는 동질성 연구를 수행하고, 이 같은 연구 결과와 IBT 양립성 원칙을 토대로, 표 2 의 문항 구성 요소, 즉, 입력과 응답 방식, 입력과 응답시간 통제, 문항의 제시 순서, 문항 간 이동 여부, 답안 수정 여부 등을 결정할

[7] 환경 타당도의 구성 요소에는 시험 주제, 과업 상황, 과업 순서, 시간 제한, 지식, 입력과 출력, 과업 요구 사항, 시험 시행 조건 등이 있다.

수 있을 것이다. 또한 제시 속도 및 응답 속도에 대한 환경 타당성 검증을 토대로 듣기 및 읽기의 적정 속도(예, 180, 200wpm)도 결정해야 할 것이다. 마지막으로 NEAT 는 한국 영어교육 사상 최초로 IBT 로 시행되는 만큼 지역·계층 간 시험 수행의 차이를 최소화하여 시험의 공정성과 형평성을 확보할 수 있도록 시험 체제를 결정해야 할 것이다[8].

Ⅳ. 문항 유형 확정 모델(Item-Types Decision-Making Model)

문항 유형 확정 모델(Item-Types Decision-Making Model)은 검사지 구성의 절차와 방법에 관한 일종의 흐름도(flowchart)로서 문항 유형 풀, 문항 유형 타당성 평가 기준과 절차, 문항 유형 은행, 모듈형(modular) 평가 목표 이원분류표, 상보형(complementary) 평가 목표 이원분류표, 예비 시행, 예비 시행 결과에 따른 평가 목표 이원분류표의 수정·보완 등으로 구성된다. 문항 유형 확정 모델의 작동 과정을 간략하게 살펴보자. 문항 유형 확정 모델의 입력부에 해당하는 문항 유형 풀에 기존 표준화 시험의 문항이나 IT 기술과 융합된 신 유형의 문항을 입력한다. 다음 문항 유형 타당성 평가 기준과 절차에 따라 입력된 각 문항 유형의 타당성 및 특성을 평가하여 문항 유형 은행에 저장 여부를 결정한다. 저장이 결정되면, 모듈형 이원분류표의 내용 영역 체계, 행동 영역 체계 및 문항 유형 특성 체계의 해당란에 이를 등재한다. 모듈형 이원분류표의 내용 영역 체계, 행동 영역 체계 및 문항 유형 특성 체계로부터 매 시행 시기 마다 수준별(2·3 급), 기능별(듣기·읽기) 시험 간에 평가 내용과 평가 요소(문항 유형)가 서로 상보적 관계에 있도록 상보형 평가 목표 이원분류표를 구성한다. 이를 토대로 각 시험의 문항지를 제작하여 시험을 시행한다. 시행 후, 시행 결과 분석을 토대로 시행된 각 문항의 타당성을 실증적으로 검증하여, 수정·보완한다. 다음에서는 문항 유형 확정 모델의 각 구성 요소를 보다 구체적으로 살펴보고자 한다.

---

[8] 수능 듣기 횟수에 따라 지역별 편차가 생기면 시험의 공정성 문제를 야기하기 때문에 4 차 실험평가에서 듣기를 한번 들려 줄 때와 두 번 들려 줄 때의 이해도의 차이를 비교하는 실험을 했으며, 이 실험 결과에 따르면, 서울과 대도시 학생은 한 번 들려줄 때보다 두 번 들려줄 때 약 8-10 점 정도 상승했지만, 중소도시나 읍/면 지역 학생은 0.3-3 점 정도 상승하는데 그쳤으며, 이에 따라 수능 듣기는 한번 들려주는 것으로 결정되었다(한국교육과정평가원, 2005a, p. 195).

1. 문항 유형 풀(Item-Types Pool)의 구성

NEAT 의 검사지를 구성할 문항 유형을 선별하기 위해서는 먼저 문항 유형 풀(Item-Types Pool)을 구성해야 한다. 문항 유형 풀은, 표 3 에서 보는 바와 같이, 언어능력의 구성소 및 기능별로 문항을 범주화한 것을 말한다. 문항 유형 풀에 포함될 표준화 시험은 앞서 언급했던 시험 체제 분석에 포함된 국·내외 주요 수행력 시험(TOEFL, TOEIC, IELTS, EIKEN, CET, TEPS, FLEX, MATE 등)과 동북아 3 국의 입학시험(수능, 일본입시센터시험, 중국고시) 등이다. 또한 영어교육 전문가가 구성한 신 유형의 문항과 IT 기술을 접목하여 개발된 문항도 이 문항 유형 풀에 포함될 수 있다. 이 같은 표준화 시험의 문항이나 신 유형의 문항을, 표 3 에서와 같이, 문법, 어휘, 듣기 단일 문항, 듣기 복합 문항, 읽기 단일 문항, 읽기 복합 문항 등으로 범주화하여 문항 유형 풀에 등재한다.

표 3

문항 유형 풀(Item-Types Pool)

언어 능력 구성소 및 기능	문항 유형 풀(Item-Types Pool)
문법 문항	$\{G_1, G_2 \cdots \cdots \cdots \cdots \cdots \cdots \cdots G_{n-1}, G_n\}$
어휘 문항	$\{V_1, V_2 \cdots \cdots \cdots \cdots \cdots \cdots \cdots V_{n-1}, V_n\}$
듣기 단일 문항	$\{L_1, L_2 \cdots \cdots \cdots \cdots \cdots \cdots \cdots L_{n-1}, L_n\}$
듣기 복합 문항(1 대화문 2 문항)	$\{Lc_1, Lc_2 \cdots \cdots \cdots \cdots \cdots Lc_{n-1}, Lc_n\}$
읽기 단일 문항	$\{R_1, R_2 \cdots \cdots \cdots \cdots \cdots \cdots \cdots R_{n-1}, R_n\}$
읽기 복합 문항(1 지문 2/3 문항)	$\{Rc_1, Rc_2 \cdots \cdots \cdots \cdots \cdots \cdots Rc_{n-1}, Rc_n\}$
신 유형 문항	$\{Ni_1, Ni_2 \cdots \cdots \cdots \cdots \cdots \cdots Ni_{n-1}, Ni_n\}$

<범례>: G: Grammar, V: Vocabulary, L: Listening, R: Reading, N: New Item

2. 문항 유형 타당성 평가 기준 및 점수화 절차

문항 유형 풀에 등재된 문항 유형의 타당성을 평가하기 위해, 문항 타당성 평가 요소와 평가 기준을 설정해야 한다. 문항 유형 결정 원리(김용명, 2010)에서 문항 유형의 질적 타당도를 평가하기 위한 5 기준, 즉, 연계성, 실제성, 상호작용성, IBT 양립성, 환류 효과성과 그 하위 기준 및 평가 요소를 설정한다. 이 같이 설정된 평가의 기준과 하위 기준 및 평가 요소에 따라 문항 평가자는 문항 유형 풀에 등재된 문항 유형의 타당도를 Likert 5 점 척도로 평가한다. 한편 검사지 구성 원칙(김용명, 2010)에서 문항 유형 특성 평가를 위한 3 기준, 즉, 통합성, 주축성, 위계성과 그에 따른 평가

요소를 설정한다(다음 표 4 참조). 이 같이 구성된 문항 특성 평가 기준과 평가 요소에 따라 문항 평가자는 등재된 문항 유형의 특성을 평가한다.

1) 문항 유형 타당성 평가 기준

(1) 연계성(Connection)

연계성은 시험 과업이 교육과정의 목표, 내용, 성취 기준과 연계된 정도를 말한다. 이에 대한 평가 요소는 "문항 유형이 교육과정의 성취 기준과 연계되어 있는가?" 로 조작적 정의를 한다.

(2) 실제성(Authenticity)

실제성은 세 하위 기준, 즉 'TLU 영역', '필요 불가결성', '이해 방식' 으로 구성된다. 'TLU 영역' 은 실제성 원리의 핵심 요소로 시험 과업의 특성과 TLU 과업의 특성 간의 일치 정도를 말한다. 따라서 이 기준에 대한 평가 요소는 "문항 유형의 특성이 TLU 과업의 특성과 일치 하는가?" 로 조작적 정의를 한다.. '필요 불가결성' 은 시험 과업이 해당 TLU 영역에서 필요 불가결한 정도를 말한다. 따라서 이에 대한 평가 요소는 "문항 유형이 TLU 영역에 비춰보아 필요 불가결한 것인가?" 로 정의한다. '이해 방식' 은 실제 언어사용 상황에서는 텍스트에 따라 읽기 · 듣기의 이해 방식을(상향식, 하향식, 상호작용식) 선별적으로 활용하는 것을 말한다. 따라서 이에 대한 평가 요소는 "문항 유형이 TLU 상황에서 듣기 · 읽기의 이해 방식을 반영하고 있는가?" 로 정의한다.

(3) 상호작용성(Interactiveness)

상호작용성은 두 하위 요소, 즉, '상호작용성', '전략적 능력' 로 구성된다. '상호작용성' 에 대한 평가 요소는 "문항 유형이 수험자의 다양한 언어능력과 상호작용하는가?" 로 조작적 정의를 한다. '전략적 능력' 은 실제 언어 사용 능력의 본질적 요소로서 학습자의 내재된 언어능력은 바로 이 전략적 능력을 통해 실시간 상으로 발현되고 구현된다. 따라서 이 기준에 대한 평가 요소는 "문항 유형이 실시간 상으로 수험자의 전략적 능력과 관계하는가?" 로 조작적 정의를 한다.

(4) IBT 양립성(Compatibility)

IBT 양립성 기준은 '양립성', '친화성', '상호작용성' 등 세 하위 기준으로 구성된다. '양립성' 에 대한 평가 요소는 "문항 유형이 IBT 로 구현이 용이한가?", '친화성' 은 "IBT 문항 유형이 수험자 친화적으로 운영될 수 있는가?", '상호작용성' 은 "IBT 문항 유형이 수험자의 언어능력과 상호작용하는가?" 로 각각 조작적 정의를 한다.

표 4

문항 유형 타당성 평가 기준 및 평가 요소

영역	하위 영역	평가 요소	1 2 3 4 5
연계성	성취 기준	문항 유형이 교육과정의 내용 및 성취기준과 연계성이 있는가?	
실제성	TLU 영역	문항 유형의 특성이 TLU 과업의 특성과 일치하는가?	
	필요 불가결성	문항 유형이 TLU 영역에 비춰보아 필요 불가결한 것인가?	
	이해 방식	문항 유형이 TLU 상황에서 듣기/읽기 이해 방식을 반영하는가?	
상호작용성	상호 작용성	문항 유형이 수험자의 다양한 언어능력과 상호작용하는가?	
	전략적 능력	문항 유형이 실시간 상으로 수험자의 전략적 능력과 관계하는가?	
IBT 양립성	양립성	문항 유형이 IBT 로 구현이 용이한가?	
	친화성	IBT 문항 유형이 수험자 친화성을 지니고 있는가?	
	상호 작용성	IBT 문항 유형이 수험자의 언어능력과 상호작용을 촉진시키는가?	
환류 효과성	교수 /학습	문항 유형이 교수 및 수업 활동에 긍정적 영향을 주는가?	
	교실 평가	문항 유형이 교실 평가에 긍정적 영향을 주는가?	
통합성	통합성	문항 유형이 독립형 문항 또는 기능 통합형 문항인가?	독립형:통합형
주축성	주축성	문항 유형이 주축성 문항 또는 주변성 문항인가?	주축성:주변성
위계성	위계성	문항 유형에 대한 문항 반응 곡선이 L 형, M 형, 또는 H 형인가?	L 형-M 형-H 형

(5) 환류 효과성(Washback Effectiveness)

환류 효과성은 '교수·학습', '교육 평가' 등 두 하위 기준으로 구성된다. '교수·학습'은 "문항 유형이 교수 및 수업 활동에 긍정적 영향을 주는가?", '교육 평가'는 "문항 유형이 교실 평가에 긍정적 영향을 주는가?"로 각각 조작적 정의를 한다.

(6) 통합성(Integration)

통합성은 시험 과업이 적어도 2 개 이상의 언어 기능이 통합된 것을 말한다. 따라서 이에 대한 평가 요소는 "문항 유형이 기능 독립형 문항 또는 통합형(연계형) 문항인가?"로 정의를 한다.

(7) 주축성(Pivotality)

주축성은 시험 과업이 주축의 역할을 하는 정도를 말한다. 따라서 이의 평가 요소는 “문항 유형이 주축성 문항 또는 주변성 문항인가?”로 정의를 한다.

(8) 위계성(Hierarchicality)

위계성은 시험 과업의 복잡도 또는 난이도와 수험자의 언어능력 수준(상·중·하)과 일치하는 정도를 말한다. 따라서 이에 대한 평가 요소는 “문항 유형에 따른 수험자의 문항 반응 곡선이 L 형, M 형, H 형 중, 어느 유형에 속하는가?”로 정의를 한다.

2) 문항 유형 타당성 평가 절차 및 점수화 방법

문항의 타당도 평가를 위한 8 개의 평가 기준과 그에 따른 하위 기준 및 평가 요소에 따라 문항 유형 풀에 등재된 각 문항 유형의 타당성 및 문항 유형의 특성을 평가한다. 문항 유형 타당도 및 특성 평가의 구체적 방법과 점수화 절차는 다음과 같다.

(a) 문항 평가자는 표 4 의 8 개의 평가 기준, 하위 기준 및 평가 요소에 따라 문항 유형 풀에 등재된 각 문항 유형을 Likert 5 점 척도로 평가한다.
(b) 문항 유형 타당성과 관련된 기준인 연계성, 실제성, 상호작용성, IBT 양립성, 환류 효과성 등은 1 점에서 5 점까지 척도 점수를 부여하지만, 문항 유형 특성과 관련된 기준인 통합성, 주축성, 위계성은 제시된 해당 특성을 선택한다.
(c) 점수화 방법은 비보상 통합 점수(non-compensatory composite scores) 체계로 하며, 분할 점수(cut-off score)는 4 점으로 한다. 따라서 평가 요소 중, 어느 한 요소에서라도 4 점 미만을 받으면, 해당 문항 유형은 문항 유형 은행에 저장될 수 없다.
(d) 모든 평가 요소에서 4 점 이상을 받고 동시에 평균 점수가 4 점 이상일 경우, 문항 유형 은행에 저장될 수 있다[9].
(e) 문항 유형 특성과 관련된 통합성, 주축성, 위계성은, 필요한 경우, 수능 외국어(영어) 영역의 시험 결과(2005-2010)와 NEAT 예비 시험 결과를 참조하여 평가할 수 있다.

[9] 이와 같이 비보상 통합 점수 체계에 따라 엄격한 기준을 설정한 이유는 NEAT는 전국적으로 시행되는 고부담 시험이며, 따라서 문항 유형 자체는 거의 완벽해야 시험의 공정성과 항상성을 담보할 수 있기 때문이다.

이상에서 논의한 8 개의 문항 타당성 평가 기준은 문항 개발의 전 단계에 걸쳐 일종의 평가 규준 역할을 할 것이다. 즉, 문항 유형 개발 단계에서는 문항 개발의 지침 역할을 할 것이며, 문항 타당성 평가 단계에서는 평가 기준 역할을 할 것이며, 시행 후에는 문항 타당도 검증의 기준으로 활용할 수 있을 것이다.

3. 문항 유형 은행(Item-Types Bank) 구성

문항 타당성 평가 기준에 따라 검증된 문항을 유형화 또는 범주화하기 위한 수단으로서 '문항 유형 은행(Item-Types Bank)'을 구성한다. 문항 유형 은행은, 그림 1 과 그림 2 에서 보는 바와 같이, 내용 영역 체계, 행동 영역 체계, 문항 유형 특성 체계 등 세 구성소로 구성되며, NEAT 의 검사지에 포함될 수 있는 이용 가능한 모든 문항 유형을 내용 및 행동 영역에 따라 범주화한 것을 말한다. 표 5 와 표 6 에서 추론할 수 있듯이, 문항 유형 은행은 복수 시행되는 NEAT 의 문제 은행 시스템의 이론적 틀의 역할을 할 뿐만 아니라 평가 목표 이원분류표 구성의 모체(母體) 역할을 한다(다음 절 참조).

그림 1

NEAT 듣기 시험이 문항 유형 은행(Item-Types Bank)

문항 유형 구성소 / 문항 유형	내용 영역 체계									행동 영역 체계							문항 유형 특성 체계						
	대화(Dialogue)					독화(Monologue)						이해, 적용 및 사고력					통합성		주축성		위계성		
	가정 챔플	학교 챔플	사회 챔플	직장 챔플	문화 챔플	실용 독화	뉴스	감의	감연	어휘	문법	사실적	적용력	추론적	종합적	창의적	단일형	통합형	주축형	주변형	L 형	M 형	H 형
IP_1																							
IP_2																							
IP_3																							
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
IP_{n-1}																							
IP_n																							

그림 2

NEAT 읽기 시험이 문항 유형 은행(Item-Types Bank)

문항 유형 구성소 / 문항 유형	내용 영역 체계									행동 영역 체계							문항 유형 특성 체계						
	대화(Dialogue)					독화(Monologue)						이해, 적용 및 사고력					통합성		주죽성		위계성		
	가정샘플	학교샘플	사회샘플	직장샘플	문화샘플	실용독화	뉴스	감의	감연	어휘	문법	사실적	적용력	추론적	종합적	창의적	단일형	통합형	주죽형	주변형	L형	M형	H형
P_1																							
P_2																							
P_3																							
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
P_{n-1}																							
P_n																							

4. 고정형(Fixed) 평가 목표 이원분류표

고정형(fixed) 평가 목표 이원분류표는 현행 수능 외국어영역에서처럼 내용 영역과 행동 영역이 매 시행되는 시험마다 변화 없이 고정되어 있고, 많은 경우, 내용 영역과 행동 영역의 조합에 따른 문항 유형까지 정해져 있는 이원분류표를 말한다. 이 같은 고정형 평가 목표 분류표는 현행 수능 체제와 같이 단일 수준의 단일 시행 체제에서는 시험의 항상성과 동일성을 유지하기가 비교적 용이하고, 출제 과정도 비교적 안정적으로 운영할 수 있는 장점이 있다. 그러나 2014 수능 체제와 NEAT 와 같이 수준별(A·B 형 또는 2·3 급), 기능별(듣기·읽기·말하기·쓰기) 복수 시행 체제에서는 매 시행되는 수준별, 기능별 시험 간에 상보성을 파악하기가 어렵고, 각 시험의 평가 요소 간의 간섭 효과가 있을 수 있다.

현행 수능 평가 목표 이원분류표(한국교육과정평가원, 2010c)는 이 같은 한계점뿐만 아니라 구성상에도 많은 문제점을 내포하고 있다. 그림 3 에서 보는 바와 같이, 현행 외국어(영어)영역의 듣기 시험의 경우, 내용 영역 체계가 대화와 담화로만 구성되어있을 뿐 더 이상의 하위 영역이 설정되어 있지 않다. 따라서 수능 듣기 17 문항을 대화와 담화로 분류하는 역할만 할 뿐이며, 문항의 실제성의 정도나 각 문항들이 서로 상보적 분포를 이루고 있는지 파악하기 어렵다. 또한 읽기 시험의 내용 영역은 글의 종류에 따라 논설문, 설명문, 문학, 실용문, 기타로 구성되어 있다. 이처럼 글의 종류에 따라 내용 영역을 구성할 경우, 다음과 같은 평가 이론상의 오류와 출제상의 문제점이 있다.

그림 3

현행 수능 외국어(영어)영역의 평가목표 이원분류표

내용 영역 \ 행동 영역			어휘	문법성 판단력	이해			적용	문항수	비율 (%)
					사실적	추론적	종합적			
이해 기능	듣기	대화								
		담화								
	읽기	논설문								
		설명문								
		문학								
		실용문								
		기타								
표현 기능	말하기	대화								
		담화								
	쓰기	문장								
		문단								
문 항 수									50	
비율(%)										100

먼저 읽기 이론(김용명, 1991)에 따르면, 글의 종류는 형식 스키마(formal schema)에 속하며, 이는 읽기 능력의 주요 직접적 평가 요소에 속한다. 따라서 수능 외국어(영어) 시험은 형식 스키마라는 직접 평가 요소를 내용 영역 기준으로 삼는 논리적 오류에 빠져 있으며, 그 결과 문항 개발 시, 형식 스키마의 이해 능력과 관련된 문체나 글의 구조에 관한 문항을 출제할 수 없다는 한계점이 있다. 또한 출제 과정에서 글의 종류에 따라 내용 타당도를 따져야 하므로 해당 지문의 상보성을 파악하기가 쉽지 않다. 실제 출제 과정에서 글의 종류에 따라 출제를 한 후, 소재나 학문 영역에 따라 내용 타당도를 재분석해보면, 출제된 문항이 특정 소재나 특정 학문 영역에 치우친 경우가 많다. 이때 소재 겸침으로 인해 이미 출제된 문항을 내리고 재 출제하는 경우가 종종 발생하기도 한다. 보다 본질적인 문제는 글의 종류에 따른 분류는 상보성 원칙에 위배되는 경우가 많아서 수험자의 독서 성향이 시험 점수에 영향을 미칠 수 있다는 점이다. 이는 시험의 공정성 및 형평성의 문제를 야기할 수 있다.

5. 모듈형(Modular) 평가 목표 이원분류표

문항 유형 은행에서 문항 유형을 추출하여 검사지를 구성하기 위해서는 NEAT 의 평가 목표, 평가 요소, 출제 및 시행 체제에 부합하는 평가 목표 이원분류표를 구성해야 한다. 앞서 논의한 바와 같이, NEAT 는 수준별(2·3 급)로 복수 시행 체제에 따라 IBT 방식으로 시행된다. 이 같은 시험 및 시행 체제에 부합하기 위해서는, NEAT 는 매 시행되는 수준별,

기능별 시험 간 상보성 원칙에 따라 '등거리성'을 확보하여 시험 간 평가 요소의 간섭 효과를 최소화해야 할 것이다. 이와 같이 시험 간 간섭 효과를 극소화하기 위해서는 NEAT 는 현행 수능과 같은 고정형 평가 목표 이원분류표는 보다는 '모듈형(modular)' 평가 목표 이원분류표가 더 타당할 것으로 생각한다. 모듈형 평가 목표 이원분류표는, 다음 그림 4 와 그림 5 에서 보는 바와 같이, 실제성 원리에 따라 평가 목표 이원분류표의 각 영역과 그 하위 영역의 구성소를 가능한 한 세분하여 구성하지만, 상보성 원칙에 따라 각 영역과 그 하위 영역 간에 서로 상보적 분포를 이루도록 구성한 것을 말한다. 모듈형 평가 목표 이원분류표의 구성소는 내용 영역 체계, 행동 영역 체계, 문항 유형 특성 체계로 구성된다. 고정형 이원분류표와 달리 문항 유형 특성 체계를 따로 설정한 논리적 근거는 NEAT 는 연간 수회에 걸쳐 복수로 시행되는 고부담 시험이므로 시험의 항상성과 동등성이 그 어떠한 시험보다 중요하며, 또한 이 같은 시험의 안정성과 동등성은 바로 문항 유형 특성 체계의 구성 요소 간의 적절한 균형을 통해서 확보할 수 있기 때문이다. 다음에서는 모듈형 평가 목표 이원분류표의 각 구성소를 구체적으로 살펴보고자 한다.

1) 모듈형(Modular) 평가 목표 이원분류표의 내용 영역 체계

(1) 듣기 시험의 내용 영역 체계

현행 수능 외국어(영어) 영역의 듣기 내용 영역 체계의 문제점을 보완하고 동시에 NEAT 의 평가 목표, 평가 요소에 부합할 수 있도록 NEAT 의 듣기 시험의 내용 영역 체계를 연계성, 실제성, 상보성, 위계성 등의 원칙에 따라 그림 4 와 같이 모듈형으로 구성한다.

먼저 Nunan(1991)의 구두 언어 형식 분류에 따라 듣기 시험의 내용 영역을 대화(dialogue)와 독화(monologue)로 분류한다. 대화는 대화 상황과 주제에 따라 대화의 구조가 결정되며, 의사소통 기능을 통해 의사소통 목적이 실현된다. 따라서 대화의 하위 내용 영역은 2007 개정 교육과정의 성취기준의 연계성 논리[10] 와 Estaire 와 Zanon(1994)가 제안한 주제 생성자(thematic generator)의 근접 동심원 원리에 따라[11], 그림 4 에서 보는 바와 같이, 대화 상황 및 주제, 의사소통 기능으로 구성한다. 대화 상황 및 주제의 하위 영역은 일상생활 관련 상황과 주제(예, 의식주, 가정생활 등), 학교생활 관련 상황과 주제(예, 수업, 학습 활동, 시험 등),

[10] 개정 교육과정에서는 '가까운 주변의 화제(자신, 가족, 학교, 등)' → '일상적 화제' → '(친숙한) 일반적 주제' → '일반적 주제' → '다양한 주제' 로 단계적이고 체계적인 확대 적용으로 성취기준의 연계성을 고려하였다(김진석, 2009, p. 61).

[11] 주제 생성자(thematic generator)는 주제나 화제의 근원성(closeness/remoteness)의 정도에 따라 안쪽 동심원부터 자신→가정→학교생활(교사, 급우 등)→자기 주위 세상→상상의 세계로 구성되어 있다(Estaire 와 Zanon, 1994, p. 21).

사회생활 관련 상황과 주제(관공서, 상점, 교통, 통신 등), 직장생활 관련 상황과 주제(예, 취업, 승진, 업무 활동 등), 문화생활 관련 상황과 주제(취미, 오락, 미디어, 강연, 강좌 등)로 구성한다.

그림 4

듣기 시험의 모듈형(modular) 평가목표 이원분류표

행동 영역 처겨 / 내용 영역 처겨				이해, 적용 및 사고력				문항 유형 특성 처겨							문항수	출제빈도
								통합성		주죽성		위계성				
				사실적	적용력	추론적	종합적	독립형	단일형	주죽형	주변형	L형	M형	H형		
듣기	대화	가정생활	친교 활동													
			사칠적 정보 교판													
			저적 태도 표현													
			감정적 표현													
			도덕적 표현													
			설득적 표현													
			문제 해결													
		학교생활	친교 활동													
			사칠적 정보 교판													
			저적 태도 표현													
			감정적 표현													
			도덕적 표현													
			설득적 표현													
			문제 해결													
		사회생활	친교 활동													
			사칠적 정보 교판													
			저적 태도 표현													
			감정적 표현													
			도덕적 표현													
			설득적 표현													
			문제 해결													
		문화생활	친교 활동													
			사칠적 정보 교판													
			저적 태도 표현													
			감정적 표현													
			도덕적 표현													
			설득적 표현													
			문제 해결													
	독화	실용독화	안내 독화(방송)													
			광고 독화													
			제안 독화													
			합의 독화													
		강의	문학													
			인문학													
			사회과학													
			자연과학													
			기술공학													
			판경학/의학													
		강연	정치/경제													
			사회/문화													
			건강/보건													
			예술/스포츠													
			일상생활													
		미디어	방송 뉴스													
			신문/잡지													
			인터넷													
문항 수																
출제빈도																

또한 각 주제나 상황의 하위 영역은 교육과정과의 연계성을 고려하여 개정 교육과정의 의사소통 기능의 7 개 범주에 따라 친교 활동(인사, 소개 등), 사실적 정보 교환(사실적 정보, 사실 묘사, 경험 등), 지적 태도 표현(동의, 제안, 확신, 의무 등), 감정 표현(희로애락, 불평 등), 도덕적 표현(사과, 변명, 후회 등), 설득과 권고(설득, 요청, 충고 등), 문제 해결(원인·결과, 이해 점검, 전화하기 및 받기 등)로 구성한다. 한편 독화의 하위 내용 영역은 실용 독화(experiential monologue), 강의, 강연, 미디어로 구성한다. 실용 독화엔 안내 독화(방송), 광고 독화, 항의 독화 등이 있으며, 미디어에는 방송, 신문·잡지, 인터넷 등이 있다.

이상과 같이 듣기 시험의 내용 영역을 설정한 논리적 근거는 다음과 같다. 교수 학습 이론에 따르면(Brown, 2004, 2007; Ellis, 2003), 듣기 교육은 과업 중심 교수요목(task-based syllabus)의 지배를 받으므로 실제 듣기 수업 활동을 시험에 반영함과 동시에 교실 수업이 과업 중심 수업 활동으로 진행되도록 유도하기 위해서이다. 또한 이와 같이 주제와 상황 중심으로 듣기 내용 영역을 구성함으로써 실제성의 정도를 판단하기가 용이할 뿐만 아니라 상보성 원칙에 따라 각 내용 영역 및 그 하위 영역이 서로 상보적 분포를 이루고 있는지 쉽게 파악할 수 있기 때문이다.

(2) 읽기 시험의 내용 영역 체계

현행 외국어(영어) 시험의 읽기 내용 영역 체계의 문제점을 보완하고 동시에 수능 언어영역[12] 및 미국의 SAT 와 ACT 의 내용 영역 체계[13] 의 분석결과를 토대로 NEAT 의 읽기 시험의 내용 영역 체계를 연계성, 실제성, 상보성 등의 원칙에 따라 그림 5 와 같이 모듈형으로 구성한다.

읽기 내용 영역의 구성 요소를 실용 담화, 문학, 인문과학, 사회과학, 자연과학, 기술공학, 예술 등으로 설정한다. 실용 담화의 하위 영역은 안내문, 광고문, 서간문, 지시문 등으로 구성한다. 문학은 수필, 이야기체 글, 소설(장편, 단편), 희곡·시나리오 등으로 구성하며, 인문학은 철학, 역사, 언어학, 인문 일반 등으로 구성한다. 사회과학은 정치학, 경제학, 사회학, 교육학, 사회 일반 등으로 구성하며, 자연과학은 물리, 화학, 생물, 지구과학, 환경학, 자연 일반 등으로 구성한다. 기술 공학은 공학, 의학, 생활 공학 등으로 구성하며, 예술은 음악, 미술, 연극, 영화, 미디어 등으로 구성한다. 이 각 하위 영역은 다시 글의 종류에 따라 설명문과 논설문으로 구성한다.

---

[12] 수능 언어영역은 문학과 비 문학으로 구분하고, 비 문학은 6 개 분야, 인문, 사회, 과학, 기술, 예술, 생활/언어로 내용 영역을 나누고 있다(한국교육과정평가원, 2010d).

[13] SAT Section 1 은 언어, 인문, 예술로, Section 2 는 인문, 소설, 과학으로, Section 3 은 사회로, ACT 는 소설, 인문학, 사회학, 자연과학으로 각각 내용 영역을 구분하고 있다(이양락 외, 2009).

그림 5

읽기 시험의 모듈형(modular) 평가목표 이원분류표

행동 영역 처겨 / 내용 영역 처겨			어휘	문법	이해, 적용 및 사고력					문항 유형 특성 처겨							문항 수	출제 빈도
										통합성		주죽성		위계성				
					사실적	적용력	추론적	종합적	창의적	단일형	통합형	주죽형	주변형	L형	M형	H형		
읽기	실용담화	만내문																
		서간문																
		광고문																
		지시문																
	문학	수필																
		이야기체 글																
		소설																
		희곡/시나리오																
	인문학	철학	설명문															
			논설문															
		언어학	설명문															
			논설문															
		역사학	설명문															
			논설문															
		인문 일반	설명문															
			논설문															
	사회과학	정치 /경제	설명문															
			논설문															
		사회학	설명문															
			논설문															
		교육학	설명문															
			논설문															
		사회 일반	설명문															
			논설문															
	자연과학	물리 /화학	설명문															
			논설문															
		생물 /지학	설명문															
			논설문															
		환경학	설명문															
			논설문															
		자연 일반	설명문															
			논설문															
	기술공학	의학 /약학	설명문															
			논설문															
		응용 공학	설명문															
			논설문															
		생물 공학	설명문															
			논설문															
	예술	음악	설명문															
			논설문															
		미술 /건죽	설명문															
			논설문															
		면극 /명화	설명문															
			논설문															
		스포츠	설명문															
			논설문															
		미디어	설명문															
			논설문															
문항 수																	?	
출제 빈도																		

이상과 같이 읽기 시험의 내용 영역을 설정한 논리적 근거는 다음과 같다. 교수 학습 이론에 따르면(Brown, 2004, 2007), 읽기 교육은 내용 중심

교수요목(content-based syllabus)의 지배를 받으므로 실제 읽기 수업 활동을 시험에 반영함과 동시에 교실 수업이 내용 중심 수업 활동으로 진행되도록 유도하기 위해서이다. 또한 이와 같이 학문 영역과 소재를 중심으로 읽기 내용 영역을 구성함으로써 실제성의 정도를 예측하기 용이할 뿐만 아니라 상보성 원칙에 따라 각 내용 영역 및 그 하위 영역이 서로 상보적 분포를 이루고 있는지 쉽게 파악할 수 있기 때문이다.

2) 모듈형(Modular) 평가 목표 이원분류표의 행동 영역 체계

Anderson 과 Krathwohl(2001)의 인지 과정 영역 목표[14], 수능 언어 영역의 행동 영역[15] 및 외국어(영어) 영역의 행동 영역[16] 등의 구성 요소의 분석 결과를 토대로 NEAT 의 듣기 및 읽기의 행동 영역 체계를 모듈형으로 구성한다(그림 4, 5 참조). NEAT 듣기 시험의 행동 영역 체계는, 그림 4 에서 보는 바와 같이, 4 구성 요소, 즉, 사실적 이해력, 적용력, 추론적 이해력, 종합적 이해력으로 구성한다. 한편 NEAT 읽기 시험의 행동 영역 체계는, 그림 5 에서 보는 바와 같이, 7 개의 구성 요소, 즉, 어휘적 판단력, 문법적 판단력, 사실적 이해력, 적용력, 추론적 이해력, 종합적 이해력, 창의적 이해력으로 구성한다.

3) 모듈형(Modular) 평가 목표 이원분류표의 문항 유형 특성 체계

문항 유형 특성 체계는, 그림 4 와 그림 5 에서 보는 바와 같이, 검사지 구성의 네 원칙, 즉, 상보성, 통합성, 주축성, 위계성으로 구성된다. 문항 유형 특성 체계는 개별 검사지의 구성, 보다 구체적으로 말하면, 모듈형 평가 목표 이원분류표로부터 상보형 평가 목표 이원분류표를 구성하는 것과 관계한다(다음 절 참조). 또한 문항 유형 특성 체계는 검사지의 질적 특성과 관계한다. 부연 설명 하면, 문항 유형 특성 체제의 구성소, 즉, 통합성, 주축성, 위계성에 따른 각 문항 유형의 적정 구성 비율을 조정함으로써 문항 특성 체계는 시험의 항상성, 다양성, 동등성, 형평성을 확보할 수 있을 뿐만 아니라 NEAT 의 수준별(2·3 급) 시험 간의 난이도 및 변별도 격차를 유지할 수 있는 일종의 지렛대로서 역할을 할 수 있다. 일반적으로 연계형 문항, 주축 문항, H 형 문항의 비율을 높이면, 시험의 항상성과 동등성을

[14] Anderson 과 Krathwohl(2001)은 Bloom(1989)의 교육 목표 분류 체계를 개정, 인지적 영역을 기억(remember), 이해(understand), 적용(apply), 분석(analyze), 평가 (evaluate), 창의(create) 등으로 구분했다.
[15] 수능 언어영역에서는 행동 영역을 어휘·어법, 사실적 이해력, 추론적 이해력, 비판적 사고, 창의적 사고 등으로 세분하고 있다(한국교육과정평가원, 2010d).
[16] 수능 외국어(영어)영역에서는 행동 영역을 어휘, 문법성 판단력, 사실적 이해력, 추론적 이해력, 종합적 이해력, 적용으로 세분하고 있다(한국교육과정평가원, 2010c).

유지하기가 용이할 것이며, 상위 변별력 확보할 수 있으므로 NEAT 의 2 급 검사지 구성에 상대적으로 더 부합할 것이다. 반면 기능 독립형 문항, 주변 문항, L 형 문항의 비율을 높이면, 시험의 다양성을 확보하기가 용이하고, 하위 학습자들의 정답률에 민감하므로 NEAT 의 3 급 검사지 구성에 상대적으로 더 적합할 수 있다. 다음에서는 제 2 장에서 논의한 바를 토대로 문항 특성 체계의 각 구성소를 살펴보고자 한다.

(1) 상보성

상보성은 문항 또는 문항 유형이 언어능력, 내용 영역 및 행동 영역의 각 구성소가 서로 상보적 분포를 이룰 수 있도록 검사지를 구성해야 한다는 원칙이다. 다시 말해, 문항 a, b, c 가 평가하고자 하는 언어능력의 구성소, 재고자 하는 내용 영역의 구성소, 측정하고자 하는 행동 영역의 구성소 각각에 대하여 상보적 관계에 있을 때 상보성 원칙을 만족시킨다고 할 수 있다. 예를 들며, 내용 일치, 빈칸 추론, 목적 찾기 문항은 사실적 이해력, 추론적 이해력, 종합적 이해력을 각각 측정하고, 또한 각각 상향식, 상호작용식, 하향식 읽기 이해 능력을 평가한다는 점에서 상보적 분포를 이루고 있으므로 상보성 원칙을 만족시킨다. 반면, 주제, 제목, 요지 추론 문항은 모두 추론적 이해력을 측정하고, 또한 모두 하향식 읽기 능력을 평가한다는 점에서 동일한 능력을 측정하고 있으므로 상보성 원칙에 위배된다. 이런 점에서 상보성은 '필수 불가결한 문항을 필요 최소한' 으로 검사지를 구성해야 한다는 원칙으로 이해할 수 있으며, 따라서 각 수준별, 기능별 검사지에 포함될 적정 문항 유형 수를 결정하는데 논리적 타당성을 제공할 수 있다. 또한 상보성은 동일 능력을 측정하는 유사한 유형은 묶어 소위 '메타 문항' 으로 분류하는 기능을 할 수 있으므로 동형 또는 평형 검사형을 구성하는데 이론적 틀을 제공해줄 수 있다.

(2) 통합성

통합성은 4 기능을 통합해서 교수해야 한다는 논리에 따라 각 기능별 시험의 독립성을 해치지 않는 범위 내에서 기능 통합형 또는 연계형 문항도 필요 최소한으로 검사지에 포함해야 한다는 원칙을 말한다. 수능 외국어(영어) 영역 문항 분석 자료(2005-2010)에 따르면, 일반적으로 연계형 문항의 정답률이 독립형 문항보다 상대적으로 낮고 정답률의 편차가 작은 경향을 보인다.

(3) 주축성

앞서 언급한 바와 같이(2 장 참조), 일반적으로 주축 문항은 시험의 안정성, 동등성을 유지하는데 기여하지만, 주변 문항은 시험의 다양성, 변화성을 주는데 기여한다. 따라서 주축성은 주축 문항을 시험의 항상성과

동등성을 확보할 수 있을 만큼 필요 최대한으로 시험의 다양성을 해치지 않을 만큼 필요 최소한으로 검사지에 포함해야 한다는 원칙을 말한다.

(4) 위계성

위계성은, 2 장에서 논의한 바와 같이, 한 검사지를 구성하는 각 문항의 난이도는 각 수험자 별(또는 각 집단 별) 수행 가능 단계와 일치할 수 있도록 위계화되어야 한다는 원칙을 말한다. 위계화의 정도를 알아보기 위해서는, 검사지에 포함 될 모든 문항의 문항 반응 곡선을 문항 유형 난이도 위계화 좌표 상에 좌표로 표시해봄으로써 각 문항이 L 형, M 형, H 형의 문항 반응 곡선 중, 어디에 속하는지를 파악할 수 있다. 이제 좌표 상에 표시된 모든 문항 반응 곡선을 문항 난이도 위계화 곡선을 따라 위계화해 봄으로써 수험자(수험자 집단)의 언어 수행력 단계(등급별)와 일치하는 문항 유형을 선별할 수 있을 뿐만 아니라 이를 통해 검사지의 총체적 난이도 및 변별도를 조정할 수 있다(김용명, 2010, pp. 389-392).

6. 상보형(Complementary) 평가 목표 이원분류표

상보형(Complementary) 평가 목표 이원분류표는 모듈형 이원분류표의 내용 영역 체계, 행동 영역 체계 및 문항 유형 특성 체계로부터 상보성 원칙에 따라 매 시행 시기 마다 수준별(2·3 급), 기능별(듣기·읽기) 시험 간에 평가 내용과 평가 요소가 서로 상보적 분포를 이룰 수 있도록(서로 겹치지 않도록) 구성한 각 개별 시험의 평가 목표 이원분류표를 말한다. 따라서 모듈형 평가 목표 이원분류표로부터 각 수준별, 각 기능별 시험에 부합하는 상보형 평가 목표 이원분류표를 구성하기 위해서는 상보성에 따른 적정 문항 수, 통합성에 따른 기능 독립형 문항과 통합형 문항의 적정 구성 비율, 주축성에 따른 주축 문항과 주변 문항의 적정 구성 비율, 위계성에 따른 L 형, M 형, H 형 문항의 적정 구성 비율 등을 선행적으로 결정해야 한다.

NEAT 의 듣기 및 읽기 시험의 적정 문항 수는 상보성 원칙에 따라 필요 불가결한 문항을 필요 최소한으로 결정될 것이지만, 2014 수능 체제 개편(안)과 듣기 문항 확대 방안(17 문항에서 25 문항) 등과 같은 정책적 측면을 고려해 잠정적으로 제안하면, NEAT 의 듣기와 읽기의 적정 문항 수는 각 25 개(가교 문항 제외)가 타당할 것으로 생각된다. 이에 대한 논리적 근거는 다음과 같다. 첫째, 시행 초기 단계에 NEAT 의 문항 개발, 출제 및 시행 부담을 줄일 수 있다. 둘째, NEAT 의 듣기와 읽기의 문항 수를 수능의 수준별 영어의 듣기와 읽기 시험의 문항 수와 연계하여 동일하게 구성함으로써 NEAT 의 수능 대체 초기 단계에 수험생의 적응을 용이하게 할 수 있어 대체의 연착륙을 기대할 수 있다.

통합성 원칙에 따라 NEAT 의 기능 독립형 문항과 기능 연계형 또는 통합형 문항의 적정 구성 비율이 결정되겠지만, 수능 외국어(영어) 영역 시험과 NEAT 의 연계성을 고려하여 잠정적으로 제안하면, 기능 독립형 문항과 기능 연계형 문항의 적정 구성 비율은 '9:1' 이 타당할 것으로 생각된다[17].

주축성 원칙에 따라 주축 문항과 주변 문항의 적정 비율이 결정되겠지만, NEAT 가 전국적으로 시행되는 고부담 시험이므로 시험의 항상성과 동등성 확보가 중요하다. 이점에 초점을 두고 잠정적으로 제안하면, 주축 문항과 주변 문항의 적정 구성 비율은 '4:1' 또는 '3:1' 가 타당한 것으로 생각된다. 전자의 경우, 항상성과 안정성을 확보하기가 상대적으로 용이하므로 NEAT 2 급에 더 부합할 것으로 생각하며, 후자의 경우, 시험에 변화성 및 다양성을 도모하기가 상대적으로 쉬우므로 NEAT 3 급에 더 적합할 것으로 생각된다. NEAT 예비 검사 결과 분석을 통해 이 제안의 타당성을 검증하여, 적정 구성 비율은 조정하여야 할 것이다.

위계성 원칙에 따라 NEAT 의 L 형, M 형과 H 형의 적정 구성 비율이 결정되겠지만, 정책적 측면과 시행의 측면을 고려하여 잠정적으로 제안하면, L 형, M 형과 H 형의 적정 구성 비율은 '3:4:3' 또는 '2:6:2' 가 타당할 것으로 생각된다. 전자의 경우, 총체적 시험의 난이도는 높을 것이며, 상·하위 학습자(수험자) 간에 변별력이 있을 것이므로 NEAT 의 2 급에 더 부합할 것으로 생각된다. 반면 후자의 경우, 총체적 난이도는 낮을 것이며, 중하위 학습자(수험자) 간에 변별력이 있을 것이므로 NEAT 3 급에 더 적합할 것으로 생각된다. NEAT 의 예비 시행 결과 분석을 통해 이 제안의 타당성을 검증, 이를 토대로 L 형, M 형과 H 형의 적정 비율은 조정할 수 있을 것이다.

NEAT 의 적정 문항 수, 통합형과 독립형 문항의 적정 구성 비율, 주축 문항과 주변 문항의 적정 구성 비율, L 형, M 형, H 형 문항의 적정 구성 비율 등이 확정되면, 이제 매 시험 마다 각 수준별, 기능별 시험에 적합한 상보형 평가 목표 이원분류표를 구성할 수 있을 것이다. 또한 매 시험에서 선별된 평가 요소를 모듈형 평가 목표 이원분류표 상의 출제 빈도란에 기록함으로써 다음 시험의 상보형 이원분류표를 구성하는데 참고할 수 있을 것이다(그림 4 와 그림 5 참조).

이상에서 논의한 바와 같이, 모듈형 평가 목표 이원분류표로부터 상보성, 통합성, 주축성, 위계성 원칙에 따라 각 수준별, 기능별 시험의 상보형 이원분류표를 구성함으로써 NEAT 의 항상성, 동등성, 변별성을 확보할 수 있을 뿐만 아니라 매 시행되는 시험 간 등거리를 유지하여 평가 요소 간의 간섭 효과를 최소화함으로써 NEAT 의 형평성과 공정성을 담보할 수 있다.

---

[17] 수능 외국어(영어) 영역 문항 분석 자료(2005-2010)에 따르면, 듣기와 말하기 연계 문항이 약 30%, 읽기와 쓰기의 연계문항이 약 20%이다. 수능은 2 기능(듣기, 읽기) 분리 시험이고, NEAT 는 4 기능(듣기, 읽기, 말하기, 쓰기) 분리 시험임을 감안하여, NEAT 의 읽기와 듣기 시험에서 연계형 또는 통합형의 적정 비율을 10%로 제안한다.

7. 상보성 평가 목표 이원분류표의 수정·보완

수준별 및 기능별 시험의 상보형 평가 목표 이원분류표가 구성되고 나면, 이를 토대로 출제 지침서를 구성한다. 출제 지침서의 주요 내용에는 각 수준별, 기능별 시험의 개념과 목표, 출제의 지침과 방향, 출제 범위(어휘 수준, 구문 수준, 담화나 대화의 길이, 지문의 길이 등), 문항 제작 지침, 문항 유형 예시 등이 포함된다.

상보형 평가 목표 이원분류표에 따라 각 수준별, 기능별 검사지를 구성하고, 이를 예비 시행한다. IBT 와 PBT 의 동질성 연구와 환경 타당도 분석을 할 수 있도록 예비 시행은 일정 실험 집단을 구성하여 IBT 와 PBT 로 동시에 실시한다. 두 집단 간의 시험 수행 결과를 비교하여, IBT 와 PBT 간의 동질성을 분석하고, 환경 타당도를 검증한다. 동질성 분석과 환경 타당도 분석 결과에 따라 상보형 평가 목표 이원분류표를 수정·보완하고, 이를 토대로 모듈형 평가 목표 이원분류표 및 시험 체제를 수정·보완 한다.

Ⅴ. 결론 및 시사점

본 연구는 NEAT 구성의 밑그림으로 제안한 문항 유형 결정 원리와 검사지 구성 원칙을 토대로 NEAT 의 문항 유형 확정과 검사지 구성의 구체적 절차와 방법에 관한 실행 설계도로서 문항 유형 확정 모델을 제안했다.

문항 유형 확정 모델은 검사지 구성에 관한 일종의 흐름도로서 문항 유형 풀, 문항 유형 타당성 평가 기준과 점수화 절차, 문항 유형 은행, 모듈형 평가 목표 이원분류표, 상보형 평가 목표 이원분류표, 예비 시행, 예비 시행 결과에 따른 평가 목표 이원분류표의 수정·보완 등으로 구성된다. NEAT 의 개발자로 가정하고, 문항 유형 확정 모델을 실행해보자. 문항 유형 확정 모델의 입력부에 해당하는 문항 유형 풀에 기존 표준화 시험의 문항이나 IT 기술과 접목된 신 유형의 문항을 입력한다. 다음 문항 유형 타당성 평가 기준과 절차에 따라 입력된 문항 유형의 타당성 및 특성을 평가하여 문항 유형 은행에 저장 여부를 결정한다. 저장이 결정되면, 모듈형 이원분류표의 내용 영역 체계, 행동 영역 체계 및 문항 유형 특성 체계의 해당란에 이를 등재한다. 모듈형 이원분류표의 내용 영역 체계, 행동 영역 체계 및 문항 유형 특성 체계로부터 매 시행 시기 마다 각 수준별(2·3 급), 각 기능별(듣기·읽기) 시험의 평가 내용과 평가 요소가 서로 상보적 관계에 있도록 상보형 평가 목표 이원분류표를 구성한다. 이를 토대로 문항지를 제작하여 시험을 시행한다. 시행 후, 시행 결과 분석을 토대로 시행된 각 문항의 타당성을 실증적으로 검증하여, 수정·보완하거나 또는 폐기한다.

NEAT 구성의 실행 설계도로 제안된 문항 유형 확정 모델은 수능 외국어(영어) 영역의 듣기 문항 확대 방안과 2014 수능 체제 개편안에도 적용가능성이 있다. 듣기 문항 확대 방안에 따르면, 듣기 문항이 현행 17 개에서 25 개로 늘어나면, 그에 따라 읽기 문항은 현행 33 개에서 25 개로 축소된다. 따라서 듣기 영역에서 어떤 유형의 문항을 늘릴 것인가와 읽기 영역에서 어떤 문항을 줄일 것인가의 문제는 문항 유형 확정 모델에 따라 결정될 수 있을 것이다. 듣기의 경우, 문항 유형 타당성 평가 기준에 따라(이 경우, 수능은 PBT 이므로 IBT 기준은 배제됨) 표준화 시험(TEPS, TOEIC 등)의 문항과 신 유형의 문항을 평가한 후, 그 결과를 토대로 듣기 영역의 25 문항을 확정할 수 있을 것이다. 한편 읽기의 경우, 상보성 원칙에 따라 현행 33 문항을 유형별로 분류한 후, 듣기와 마찬가지로 이를 문항 유형 타당성 평가 기준에 따라 평가하고, 이를 토대로 읽기 영역의 25 문항도 확정할 수 있을 것이다.

또한 듣기 문항 확대 방안과 맞물려 있는 2014 수능 체제 개편안의 수준별 영어(A·B 형)도 문항 유형 확정 모델에 따라 수준별 영어 A 형과 B 형의 검사지를 구성할 수 있을 것이다. 표준화 시험(TEPS, TOEFL 등)의 문항, 이미 개발된 NEAT 의 문항, 신 유형 문항 등을 문항 유형 타당성 평가 기준에 따라 평가한 후, 이를 토대로 모듈형 이원분류표를 구성한다. 문항 유형 특성 체계의 구성소, 즉, 상보성, 통합성, 주축성, 위계성에 따른 각 문항 유형의 적정 구성 비율을 조정함으로써 수준별 영어 A 형과 B 형에 부합하는 상보형 평가 목표 이원분류표를 구성할 수 있을 것이다. 수준별 영어 A 형의 경우, 정답률이 상대적으로 높은 독립형 문항, 주변 문항, L 형 문항의 비율을 늘리고, 정답률이 상대적으로 낮은 연계형 문항, 주축 문형, H 형 문항의 비율을 줄이는 반면, 수준별 영어 B 형의 경우, 반대로, 독립형 문항, 주변 문항, L 형 문항의 비율을 줄이고, 연계형 문항, 주축 문형, H 형 문항의 비율을 늘림으로써 수준별 영어 A 형과 B 형의 적정 난이도와 변별도 격차를 유지할 수 있을 것이다. 또한 위계성 원칙에 따라 L 형, M 형과 H 형의 적정 구성 비율을 조정함으로써 A 형과 B 형의 총체적 난이도 및 별별도를 확보할 수 있을 것이다. 마지막으로 NEAT 의 수능 대체를 대비한다면, 2014 수능 체제의 수준별 영어(A·B 형)의 모듈형 이원분류표와 NEAT(2·3 급)의 모듈형 평가 목표 이원분류표를 공통으로 구성하여, 이를 공유함으로써 대체의 연착륙을 기대할 수 있을 것이다.

또한 NEAT 의 문항 유형 확정 모델은 국가수준학업성취도평가 및 각 급 학교에서의 교실 평가에도 적용가능성이 있다. 이 경우, 시험의 성격과 시행 방식이 상이하므로 문항 유형 평가 기준을 다소 수정해야 할 것이다. 문항 유형 평가 기준에서 IBT 시행 방식을 전제로 설정한 'IBT 양립성' 기준을 삭제하고, 대신 성취도 성격에 부합할 수 있도록, 가령, '성취 기준 적합성' 을 설정하거나 기존의 '연계성' 기준의 하위 기준을 보다 세분할

수도 있을 것이다. 또한 '실제성' 은 교육과정 상의 내용이나 교과 내용으로 한정해야 할 것이며, 교실 수업에서의 교수·학습 과정을 고려하여 '환류 효과성' 기준의 하위 기준에, 가령, '교수·학습 향상성', '보정 교육 연계성' 등을 추가할 수도 있을 것이다. 또한 NEAT 및 2014 수준별 영어와 동일한 방식으로 모듈형 이원분류표를 구성할 수 있을 것이며, 문항 유형 특성 체계의 구성소의 구성 배율을 조정하여 초 6, 중 3, 고 2 에 부합하는 상보형 분류표를 작성할 수 있을 것이다. 이 경우, 교육과정에서 4 기능 통합 지도를 강조하고 있으므로 연계형 또는 통합형 문항의 비율을 늘리는 것이 타당할 것이며, 또한 위계성에 따라 저학년일수록 L 형의 비율을, 고학년일수록 H 형의 비율을 증가하는 것이 타당할 것이다.

문항 유형 확정 모델의 문제 은행 체제에 적용가능성을 논하면서 본 연구를 마무리 하고자 한다. 앞서 시뮬레이션을 해본 바와 같이, NEAT 가 연간 24 회 시행된다면, 또 2014 수능 체제에서 수준별 영어가 연 2 회 복수 시행된다면, 문제 은행식 또는 문항 공모식 출제 체제가 불가피할 것으로 예상된다. 이 같은 문제 은행식 또는 문항 공모식 출제 체제 하에서 문항 유형 확정 모델 그 자체가 문항 유형 은행 시스템의 역할을 할 수 있다. 부연 설명하면, 문항 유형 개발자가 문항을 개발하면, 이를 문항 유형 풀에 등재하고, 문항 유형 평가자가 문항 유형 평가 기준과 절차에 따라 이를 평가한다. 타당성 평가 결과가 정해진 기준에 도달하면, 문항 유형 은행에 저장되고 모듈형 평가 목표 이원분류표에 등재될 수 있지만, 정해진 기준에 미치지 못하면, 그 문항을 수정하거나 탈락시킨다.

결론적으로 말해, 문항 유형 확정 모델에 따라 구성된 모듈형 평가 목표 이원분류표는 상보형 평가 목표 이원분류표의 구성에 있어서 일종의 '평가 형판(testing template)' 역할을 할 뿐만 아니라 문항 유형 은행 구축에 있어서 일종의 '매트릭스(matrix)' 역할을 할 수 있을 것이다. 이제 본 연구에서 제안한 실행 설계도의 각 구성 부분의 세부 시행 절차도, 예를 들면, 각 수준별(2·3 급 또는 A·B 형), 기능별(듣기·읽기)의 평가 지침서, 출제 매뉴얼, 시행 세부 치침서 등을 구성할 때이다.

참고문헌

교육과학기술부.(2008a). *외국어과 교육과정(I): 교육과학기술부 고시 제 2008-160 호 [별책 14]*. 서울: 교육과학기술부.

교육과학기술부. (2008b). 국가영어능력평가시험 개발 계획 발표. *보도자료* (2008.12.18).

교육과학기술부. (2010). 국가영어능력평가시험 대입 수시에 반영 발표. *보도자료*(2010.1.7).

교육인적자원부.(2006). 인터넷 기반 국가영어능력인증시험 시행 계획 발표. *보도자료*(2006.11.3).

교육인적자원부.(2007a). *외국어과 교육과정(I): 교육인적자원부 고시 제 2007-79 호 [별책 14]*. 서울: 교육인적자원부.

교육인적자원부.(2007b). 국가영어능력평가시험 도입 기본 계획 수립 발표. *보도자료*(2007.07.30).

김낙복. (2008). 대학수학능력시험 외국어(영어) 영역의 코퍼스 언어학적 어휘 비교 분석. *영어어문교육, 14*(4), 201-221.

김용명. (1991). *상호작용 읽기 모델의 관점에서 Good/Poor Reader 의 읽기 책략 비교 연구*. 미출간 석사학위 논문, 서울대학교.

김용명. (2010). *국가영어능력평가시험(NEAT)의 문항 유형 개발과 선별 원리 및 검사지 구성 원칙*. *영어교육, 65*(4), 369-398.

김용명, 이완기, 김진석, 고현숙. (2010). *수능 외국어(영어) 영역 개선 연구*. 한국교육과정평가원, 연구보고 CAT 2010-11.

김진석. (2009). *영어과교육과정 및 평가*. 서울: 한국문화사.

성윤미. (2003). *대학수학능력시험 외국어(영어)영역의 점수 요인분석과 그 시사점*. 미출간 박사학위논문, 인하대학교, 인천.

신명신. (1999). 대학수학능력시험 영어 읽기 지문 패턴 분석. *영어교육, 54*(4), 309-326.

이경숙. (1999). 문제 수, 지문 길이, 지문 친숙도가 영어 청해와 독해시험에 미치는 영향. *영어교육, 54*(4), 327-349.

이양락, 노은희, 조윤동, 김진석, 이문복, 김용명, 박진동, 신일용, 김진구, 김영춘. (2009). *미국 SAT 와 ACT 문항 분석*. 한국교육과정평가원 연구자료 ORM 2009-7.

장경숙. (2004). 대학수학능력시험 외국어(영어)영역 읽기 난이도 예측 모형 개발. *외국어교육, 11*(1), 111-130.

중장기 대입선진화 연구회. (2010). *중장기 대입선진화 연구회-연구 발표 세미나 자료집*. 서울: 중장기 대입 선진화 연구회

한국교육과정평가원. (2005a). *수능 10 년사 I*. 서울: 한국교육과정평가원.

한국교육과정평가원. (2005b). *수능 외국어(영어) 영역 문항 분석* (비공개 자료집).

한국교육과정평가원. (2006). *수능 외국어(영어) 영역 문항 분석* (비공개 자료집).

한국교육과정평가원. (2007). *수능 외국어(영어) 영역 문항 분석* (비공개 자료집).

한국교육과정평가원. (2008). *수능 외국어(영어) 영역 문항 분석* (비공개 자료집).

한국교육과정평가원. (2009). *수능 외국어(영어) 영역 문항 분석* (비공개 자료집).

한국교육과정평가원. (2010a). *수능 외국어(영어) 영역 문항 분석* (비공개 자료집).

한국교육과정평가원. (2010b). *국가영어능력평가시험(2·급) 개발 및 운영방안*. 한국교육과정평가원, 연구보고 ORM 2010-15.

한국교육과정평가원. (2010c). *대학수학능력시험 출제 매뉴얼: 외국어(영어)영역*.

한국교육과정평가원. (2010d). *대학수학능력시험 출제 매뉴얼: 언어 영역*.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessment: A revision of Bloom's taxonomy of educational Objectives*. Pearson Education, Inc.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
Bloom, B. S. (1989). *Taxonomy of educational objectives: The classification of educational goals*. New York: Mckay.
Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.
Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. New York: Longman.
Chapelle, C. A. & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.
Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
Estaire, S., & J. Zanon. (1994). *Planning classwork: A task based approach*. Oxford: Heinemann.
Kim, Yong-Myeong. (2006). A common metric scale (CMS) on the parallel developmental sequence model. *English Teaching, 61*(4), 77-107.
Kim, Yong-Myeong. (2007a). Diagnosis and Remedy System (DRS) for teaching English on the Common Metric Scale (CMS) model. *English Teaching, 62*(2), 47-77.
Kim, Yong-Myeong. (2007b). Validaton of the common metric scale (CMS). *English Language & Literature Teaching, 14*(1), 21-44.
Nunan, D. (1991). *Language teaching methodology: A textbook for teachers*. New York: Prentice Hall.
Roever, C. (2001). *Web-based language testing. Language, learning & technology, 5*(2), 88-94.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

예시언어(Examples in): English
적용가능 언어(Applicable Language): English and Other Languages
적용가능수준(Applicable Levels): Secondary

김용명
한국교육과정평가원 대학수학능력시험연구관리본부 출제연구실
100-784 서울시 중구 정동 15-5 정동빌딩
Tel: 02-3704-3533 / C.P.: 018-267-8998
Email: Mencius@kice.re.kr / oprcms@yahoo.co.kr

Received in October 11, 2010
Reviewed in November 20, 2010
Revised version received in December 15, 2010