韓國心理學會誌 : 社會問題
Korean Journal of Psychology: Social Issues
1999, Vol. 5, No. 2, 49~58

한국인을 대상으로 한 TOEIC 타당화 연구

이 수 정

경기대학교 교양학부

본 연구에서는 우리나라 표본만을 대상으로는 타당화 작업이 시행된 적이 없는 TOEIC 검사의, 한국표본에 대한 통계적인 검증치들을 산출하였다. 3개월을 간격으로 둔 검사 재검사 신뢰도, 그리고 내적인 합치도 등의 신뢰도 지수들은 분석에 포함된 TOEIC 검사들이 상당히 내적으로 일관되며 시간의 경과에도 불구하고 점수의 안정성을 지님을 확인하여 주었다. 검사의 준거관련 타당도를 알아보기 위하여 산출된 영어의 다른 영역과의 상호관련성은 적정한 수준의 상관계수들을 산출하여 훌륭한 공인타당도를 지니는 것이 확인되었으며, 자기보고식 영어수행 점수와의 관련성은 우수한 수준의 예언 타당도를 예측하여 보게 해 준다. 99년도 2월과 3월 시험을 대상으로 이루어진 문항분석의 결과는 대부분의 문항들이 사지선다형으로서 적정한 수준의 난이도와 변별도 지수를 지니고 있음이 확인되었다.

TOEIC은 듣기, 읽기, 표현력 등의 영어 전반에 걸친 활용능력을 10점부터 990점 사이의 점수로 환산하여 영어실력을 전반적으로 평가하는 시험이다. 오늘날 TOEIC 점수는 각 개인의 영어능력을 반영하며 이를 판단하는 가장 객관적인 기준이 되고 있다.

TOEIC은 현재 세계 각국에 도입되어 널리 활용되고 있으며, 국내에서는 공공기관 및 기업체를 비롯하여 대학에서의 학점인정, 졸업자격 시험, 카투사 선발시험 등에서 유용하게 쓰이고 있다. TOEIC을 활용하는 기관 및 기업체에서는 각종 부문에 TOEIC성적을 폭넓게 적용하고 있으며, 일정 기준을 설정하여 인사고과에 반영하기도 한다. 특히 인사정책과 관련된 의사결정 시 객관적인 지표인 TOEIC 점수는 다른 자료들에 비해 상대적으로 더 중요한 근거자료로 활용되곤 하는데, 이와 같은 이유 때문에 상당수 기업의 인사담당자들은 TOEIC 성적이 과연 수험자들의 실제 언어능력을 잘 반영하고 있는가에 대하여 의문을 제기하기도 하였다. 현 연구에서는 이런 의문 사항에 응답하기 위하여 우리나라의 수험자 집단을 대상으로 TOEIC 시험이 지니는 타당성과 TOEIC 검사점수의 신뢰성에 대하여 탐색하여 보려고 한다.

인사선발 및 배치를 위하여 사용되는 도구는 무엇보다도 우선 점수의 신뢰성이나 타당성이 먼저 확인되어야 한다. 만일 시간이나 상황에 따라 응시자들의 점수가 불안정하다면 그것만큼 선발도구로서의 용도를 의심하게 하는 문제도 없을 것이다. 따라서 점수의 안정성이나 재현가능성이 검사도구의 가장 기초적인 요건이 된다. 검사의 신뢰도로

일컬어지는 이들 지표들 이외에 검사의 용도와 관련하여 중요한 문제는 검사의 타당도이다. 검사의 타당도는 검사가 측정하려는 것을 제대로 측정하고 있느냐 하는 문제와 밀접하게 관련되어 있다. 검사는 일반적으로 각 개인의 보이지 않는 추상적인 개인차들을 탐색한다. 성취도 검사들은 특정 교과목에서 일정 기간 동안 학습자들이 수행한 지적인 작업의 양을 측정하며, 성격검사는 각 개인의 내면적인 성격특질의 상대적인 양을 측정하도록 고안된다. TOEIC의 경우에는 영어에 대한 듣기와 읽기 영역에 있어서의 전반적인 숙련도의 정도를 측정한다. 따라서 타당도의 문제는 TOEIC이 얼마나 특정 영역에서의 영어 숙련도를 잘 반영하도록 고안되어 있는지 하는 문제와 관련지어져 있다.

TOEIC은 일종의 영어 적성검사이다. 적성검사는 일반적으로 해당 영역에서의 학습의 잠재력을 측정한다. 이미 수행이 일어난 성취의 정도를 측정하기보다는 특정 영역에서의 잠재가능성을 측정하여 미래에 다가올 실제적인 장면에서의 수행 정도를 예측하려는 것이다. 이 중에서도 특히 TOEIC은 영어 실무영역에서의 수행정도를 예측하여 보려는 목적으로 고안되었다. 따라서 TOEIC 시험에서 우수한 성적을 거둔 사람이라면 실제 영어 구사장면에서도 영어를 능숙하게 잘 사용할 수 있어야 한다. 이 점이 바로 TOEIC의 타당도 문제의 핵심이다. 따라서 현 연구에서는 TOEIC이 원래 제작된 목적을 그대로 잘 반영하고 있는지를 다양한 방법들을 동원하여 확인하려고 하였다.

ETS에서는 TOEIC점수와 수험자의 실제 영어능력간의 상관관계를 밝히는 대규모 조사연구를 실시하였다(Woodford, 1982). 이 상관지표는 TOEIC 시험점수의 준거관련 타당도에 대한 증거자료를 제시하여 준다. 그 결과 영어구사력 측정 인터뷰로 산출된 영어능력점수와 TOEIC의 듣기점수는 .75의 상관을, 읽기점수와는 .69의 상관을, 그리고 인터뷰 결과 산출된 영어능력 점수와 TOEIC 총점간에는 .76 정도의 상관이 존재하는 것으로 나타났다.

나아가 ETS의 연구진들(International Communication Foundation, 1998)은 TOEIC 검사의 예언타당도를 검증하여 보기 위하여 실제 상황에서의 영어수행 정도와 TOEIC점수간의 상관정도를 조사하였다. 그 결과 75개 항목으로 구성된 구체적인 행동문항들에 대한 수행 여부를 스스로 보고하게 하여 산출된 자가진단 점수와 TOEIC 점수간에 역시 긴밀한 연관성이 있음이 확인되었다. TOEIC의 읽기점수는 영어읽기가 필요한 상황에서의 수행도와 .62의 상관을 지니는 것으로 나타났으며, TOEIC의 듣기점수와 실제 듣기수행에 대한 자기평가 간에는 .65 정도의 상관관계가 존재하는 것으로 나타났다. 제 3자에게 각 응답자들의 영어실력에 대한 동일한 문항들을 토대로 평가하여 보게 한 결과 읽기, 쓰기, 말하기, 듣기, 회화 영역 모두에서 72.3% 정도의 상당히 높은 평가자간 일치도가 산출되었다.

이런 결과에도 불구하고 TOEIC을 인사선발의 도구로 사용하는 우리나라 대기업의 인사담당자들은 TOEIC시험의 점수사용에 있어서 아래와 같은 문제점들을 제기한 바 있다. 우선, 여러 번 응시하면 점수의 변동폭이 크다(실력 이외에 요령에 의해서도 점수가 향상된다)는 것이다. 둘째, TOEIC 점수는 영어 말하기 능력을 잘 예측하여 주지 못한다고 담당자들은 지적한다. 세 번째 두드러졌던 TOEIC의 문제점으로는 TOEIC이 실제 영어 수행능력을 잘 반영하지 못한다는 것이었다. 나아가 고득점자들에게 있어서의 점수의 변별력이 좋지 않다는 것이 문제점으로 지적되었다. 또한 연이어 응시하면 동일한 문제들이 반복되어 출제된다는 점과 문제의 내용이 비즈니스 쪽에만 치우쳐 있다는 점 역시 TOEIC이 당면한 문제인 것으로 지적된 바 있다.

이들 의문사항에 대한 체계적인 분석은 앞으로의 장기적인 연구들을 통하여 지속적으로 조사될

것이나, 현 논문에서는 현재 사용이 가능한 자료들을 동원하여 위의 의문사항에 응답할 수 있는 방안을 부분적으로나마 제시하여 보려고 한다.

방 법

응시자 및 자료특성

우선 신뢰도 지수를 산출하기 위하여서는 98년도 정기시험 응시자들 중에서 3개월 이내에 시험에 다시 응시한 147명의 자료를 최종적인 분석에 포함시켰다. 검사의 내적 합치도를 산출하기 위하여서는 99년도 2월과 3월 정시시험에 응시하였던 수험생 중 1000명을 무선 표집하여 그들을 대상으로 K-R 20을 산출하였다.

TOEIC시험의 준거관련 타당도 증거들을 다방면으로 수집하기 위하여서는 TOEIC 응시자들 중 말하기 시험인 SEPT와 쓰기 시험인 TWT 시험에 응시한 각기 232명과 42명의 자료를 일단 수집하였다. 이들 중에서 TOEIC을 한번 이상 응시한 사람들의 경우에는 평균점수를 각 하위 영역별 점수로 사용하였다. 실제 영어수행 정도에 대한 자기보고 점수와 TOEIC 점수간의 상관계수의 산출을 위하여서는 473명의 TOEIC 응시자들로부터 영어수행 항목들에 대한 반응이 수집되었다. 문항분석을 위하여서는 총 469명의 TOEIC 응시자들의 원점수들을 사용하였다. 부가적으로 TOEIC과 동일한 영어영역을 측정하는 검사인 경기대학교 졸업자격 영어시험에 응시하였던 253명의 수험자 자료 중 TOEIC 시험을 치른 적이 있는 78명의 자료가 또 다른 준거관련 타당도 지수의 산출을 위하여 최종적인 분석에 포함되었다.

문항분석을 위하여서는 99년도 2월과 3월 정기시험을 보았던 총 응시자들 중 1000명을 무선 표집하여 그들의 문항반응들을 분석하였다. 이를 위하여서는 고전적 검사이론과 문항반응이론이 동원되었다.

측정도구

TOEIC 시험의 타당도 연구를 위하여 포함되었던 SEPT 시험은 영어영역에 있어서 말하기 능력을 측정한다. SEPT는 5가지의 전반적인 영어 구술능력(강세, 문법, 단어, 유창성, 이해력)을 측정하기 위하여 시사영어사가 개발한 시험으로서 표준화된 점수측정방법을 도입하고 있으며 7단계로 채점된다.

TWT는 영어로 글쓰기 능력을 평가하는 영작시험이다. 전체 검사는 크게 다지선다식 50문항과, 번역 한 과제, 그리고 수필작문 한과제로 구성되어 있다. 채점은 최저 0단계부터 최고 7단계까지로 총 8단계의 등급점수들로 이루어지는데 다지선다형 문항들에서의 등급성적과 나머지 두 과제에서의 등급성적을 종합하여 최종 판단을 내린다. 이때 최종 판단은 두 사람의 훈련된 전문평정인들에 의해 이루어지는데 만일 특정 응시자의 등급을 판단하는데 있어서 전문평정인들의 의견이 불일치하는 경우에는 제 3의 평정인이 다시 채점하여 2인 이상이 일치하는 등급이 최종 성적으로 결정된다.

SEPT의 검사 재검사 신뢰도는 .78이며 내적 합치도는 .80 TWT의 재검사 신뢰도는 .68 내적 합치도는 .77인 것으로 보고된 바 있다(국제교류진흥회, 1995, 1996).

실제적인 영어수행에 있어서의 자기평가를 위하여서는 2명의 영어권 국가에 거주한 경험이 있는 영문과 교수들이 듣기, 읽기, 말하기, 쓰기 영역별로 실제상황에서 가장 영어에 대한 숙련도를 잘 반영할 것 같은 행동항목들을 일차적으로 선정하였다. 이들 중 각 숙련도별 영어 수행 수준을 7점 척도상에서 가장 잘 반영한다고 보고된 3가지씩의 항목들을 5가지 등급별로 선정하여 각 영역별로 15개씩의 항목들을 구성하였다 ("나는 미국 또는 영어권 국가에 갔을 때, 영어로 하는 간단한 인사말을 듣고 이해할 수 있다." - 듣기 5등급, "나는 미국 또는 영어권 국가에 갔을 때, 코메디 프로를 보고 내용에 포함된 유우머를 이해할 수 있다." - 듣

기 1등급). 이때 각 문항의 적합성에 대한 평가자 간 일치도는 .79로 상당히 높은 것으로 나타났다.

경기대학교 졸업영어시험은 TOEIC과 마찬가지로 영어의 듣기와 읽기 영역에서의 숙련도를 측정하도록 고안된 시험으로써 듣기시험은 60문항 읽기시험은 80문항으로 구성되어 있다. 각 하위척도에 대한 신뢰도는 듣기 .88, 읽기 .86의 K-R 20 지수를 지니는 것으로 나타났다. 문항분석 결과 평균 난이도는 듣기 .56, 읽기 .44, 평균 변별도는 듣기 .46, 읽기 .37이었다.

결 과

신뢰도

TOEIC 검사점수의 신뢰도 증거들은 크게 두 가지 종류가 탐색되었다. 점수의 안정성으로 대변될 수 있는 검사 재검사 신뢰도와 검사의 내적인 일관성을 측정하는 검사 전체의 신뢰도 지수가 바로 그것이다. 이들 자료들은 우선 TOEIC 시험을 다시 응시하면 무조건 점수가 향상된다거나 TOEIC 문제들의 내용이 일관성 없이 너무 현실적인 문제들만을 반영하고 있다든가 하는 의문에 부분적으로나마 해답을 제공하여 줄 수 있을 것이다.

재검사 신뢰도. 98년도 정기시험 응시자들 중에서 3개월 이내에 시험에 다시 응시한 147명을 대상으로 평균 상관계수를 산출한 결과, 듣기시험의 경우 점수의 안정성 계수는 .86, 읽기시험의 경우에는 .85, 그리고 총점의 경우에는 .89의 상관 정도를 지니는 것으로 나타났다. 점수의 변동폭을 계산하여 본 결과 듣기시험의 경우에는 약 11점이, 읽기시험의 경우에는 약 7점이 향상되었다. 3개월을 재검사 기간으로 하였을 때 TOEIC 총점은 약 18점 정도가 향상되는 것으로 산출되었다. 또한 일반적인 기대와는 달리 응시자들 중 많은 수(51명, 약 35%)가 TOEIC을 다시 응시하였음에도 불구하고 점수가 오히려 감소하였다. 이런 현상은 재시험이 TOEIC점수의 향상을 무조건 초래하지 만은 않는다는 사실을 반영하여 준다.

내적 일관성. 각 문항들의 내적 일관성을 산출하기 위하여서는 K-R 20을 산출하였다. 1999년도 2월과 3월의 정기시험을 대상으로 내적 일관도 점수를 산출하여 본 결과, 2월 듣기시험의 경우에는 .92, 2월 읽기시험의 경우에는 .92의 내적 일관성 점수가 산출되었다. 1999년도 3월달 정기시험에서는 듣기시험 영역에서 .91, 읽기시험 영역에서 .94의 내적 일관성 지수가 산출되었다. 이 결과는 표본 사례로 사용된 두 번의 TOEIC시험이 모두 상당히 일관성있게 개인들의 실력의 차이를 반영하여 줌을 반영하며, 나아가 원래의 더 큰 표본을 토대로 신뢰도 지수를 산출하는 경우 좀더 높은 통계치를 산출하리라 기대해 볼 수 있을 것이다.

타당도

일단 TOEIC 하위척도와 총점간의 상관계수가 산출되었다. 듣기점수와 총점간에는 평균 .95의 상관이 읽기점수와 총점간에는 평균 .94의 상관이 있는 것으로 나타났다. 듣기와 읽기 시험간의 상관은 평균 .83 정도로 적당한 정도의 수렴타당도를 지녔다.

준거관련 타당도를 알아보기 위하여 실시되었던 TOEIC의 하위 점수들과 말하기, 쓰기능력 간의 상관은 SEPT 총점과 TOEIC 총점간의 상관이 .74, 그리고 TWT 총점과 TOEIC 총점간의 상관이 .88인 것으로 나타났다. 관련 능력간의 상관이 보여주는 준거관련 타당도 지표로서의 이 정도의 상관계수는 상당히 우수한 정도인 것으로 사료된다. 즉 TOEIC 총점은 말하기 능력의 약 50%의 분산을 쓰기 능력의 약 65%의 분산을 설명하여 줄 수 있다.

표 1에는 TOEIC, SEPT, 그리고 TWT로 산출된 각 하위영역 점수들간의 상관 정도가 제시되어 있다. 듣기와 읽기능력, 그리고 듣기와 쓰기능력 간에는 .80을 넘는 높은 정도의 상관이 있었으나 말

표 1. 하위검사들간의 상관계수

	듣 기	읽 기	말하기	쓰 기
듣 기				
읽 기	.83(274)			
말하기	.76(232)	.64(232)		
쓰 기	.86(42)	.77(42)		

( ) 안은 사례 수

하기 능력과 다른 영어능력들 간에는 상대적으로 더 낮은 관련성이 있었다. 이 점은 아마도 시험의 양식이 서로 달랐음에 기인했을 것인 바, 모두 지필검사의 양식으로 치루어졌던 TOEIC이나 TWT에 비해서 구두시험의 양식이 지배적이었던 SEPT가 지니는 검사양식 상의 독특성에 기인했기 때문인 것 같았다.

TOEIC의 예언타당도라고도 할 수 있을 실제상황에서의 영어 구사력에 대한 자기평가 점수와 TOEIC 점수들간의 상관은 듣기점수와 수행점수와의 상관이 .56, 읽기점수와 수행점수와의 상관이 .57 그리고 TOEIC 총점과 수행총점간의 상관이 .59인 것으로 나타났다. 일반적으로 적성검사의 예언타당도가 .30이상을 획득하기가 어려운 점을 고려한다면 현 연구결과 산출된 예언타당도 지수로서의 수행점수와 TOEIC 총점간의 상관은 비교적 우수한 것이라 결론지을 수 있다.

교과 영역 별로 좀더 세분화된 공인타당도와 변별타당도 지수를 얻기 위하여, 동일한 영어 영역을 측정했던 경기대학교의 졸업자격시험과 TOEIC 시험의 하위영역 점수들을 토대로 MTMM(다특질 다방법)분석을 시행하였다. 표 2에는 MTMM 결과가 요약되어 있다. 대각행렬에는 각 하위 척도의 신뢰도 지수가 산출되어 있으며 나머지 칸에는 수렴타당도와 변별타당도 지수들이 산출되어 있다. 첨자가 동일한 경우 그들은 같은 종류의 타당도 지수를 나타내는데, 예컨데 첨자 b의 경우는 동일한 교과영역을 서로 다른 시험방식(monotrait - heteromethod)으로 측정한 지표를 보여 준다. 이들 상관계수들은 수렴타당도에 대한 지표로서 각 척도에서의 신뢰도계수를 제외한, 동일한 행, 렬의 어느 수보다 더 큰 관련성을 지녀야 한다. 동일한 내용을 측정하는 b 값들에 비해 동일한 검사방식으로 서로 다른 교과영역을 측정(heterotrait - monomethod)하는 첨자 c 값들은 수렴타당도 지수들보다는 상관의 정도가 낮아야 하며 서로 다른 영역을 서로 다른 방식으로 측정(heterotrait - heteromethod)하는 d 값들보다는 높아야 한다. c와 d에 해당하는 상관지표들은 검사의 변별타당도 지수를 측정한다.

표 2. 경기대 시험과 TOEIC 시험에 대한 MTMM 분석

	경기듣기	경기읽기	TOEIC듣기	TOEIC읽기
경기듣기	$.88^{a}$			
경기읽기	$.67^{c}$	$.86^{a}$		
TOEIC듣기	$.76^{b}$	$.62^{d}$	$.92^{a}$	
TOEIC읽기	$.43^{d}$	$.82^{b}$	$.83^{c}$	$.93^{a}$

경기대 졸업자격시험과 TOEIC 시험의 경우에는 같은 영역을 측정하는 듣기시험간 읽기 시험간 상관이 그 어느 상관계수들보다 높아야 했음에도 불구하고 TOEIC 시험의 경우 경기대 시험보다 하위척도간 상관이 상대적으로 좀 높은 것으로 나타났다. 따라서 읽기시험 간 수렴타당도는 TOEIC 시험의 하위 척도간 상관보다 약간 낮았다. 이런 경우 TOEIC의 하위척도에 포함되었던 문제의 유형을 좀더 그 척도의 특성을 민감하게 반영하도록 구성한다면 쉽게 개선될 문제라 사료되었다. 그러나 듣기시험의 경우에 있어서는 TOEIC과 경기대 시험간의 상관이 .76, 즉 상당히 낮은 정도의 수렴타당도를 지니는 것으로 나타났다. 이 같은 사실은 두 가지 가능성을 시사하는데 상관을 산출한 하위척도들의 신뢰도가 낮거나 표본의 특성을 반영하여 주기 때문이다. 경기대학교 졸업시험의 경우 새로이 제작된 시험이었기에 검사의 통계적인 특질에 대한 좀더 많은 연구가 이루어져야만 이에 대

한 정확한 답을 얻을 수 있으리라 생각된다. 그러나 다만 이 자료로 관찰할 수 있는 결론은 TOEIC 시험보다는 경기대 시험이 신뢰도지수가 상대적으로 낮았다는 사실이다. 따라서 경기대 시험이 좀더 견고한 검사 통계치를 지니게 되고 TOEIC 하위검사들의 각 영역에 대한 특수성이 증가한다면 이들 간에 산출되는 검사 통계치들은 훨씬 우수해질 것이라 추정하여 볼 수 있다.

문항분석

문항의 특질을 분석한다는 것은 검사 전체의 질을 판단하는 데에 무척이나 중요한 요인이 된다. 검사의 총점이 신뢰롭기 위하여서는 그 검사에 포함된 각각의 문항들이 나름대로 훌륭한 문항 특질들을 유지하고 있어야 한다. 예컨대 문항들이 너무 어렵기만 하다든가 너무 쉽기만 하다면 응시자들의 실력에 의해서가 아니라 상황이나 응시자들의 특성들에 의하여 점수가 우연히 변동될 확률이 커지게 된다. 단일 시험이 무척 어려워서 거의 아무도 정답을 알아내기가 힘들다면 각 응시자들이 몇 개의 문항들을 우연히 맞추었느냐가 총점을 결정하는 데에 지대한 영향을 미칠 것이다.

문항의 변별력 역시 점수의 신뢰도를 결정하는 데에 중요한 요인이 될 수 있다. 문항의 변별력이 높다는 사실은 각 문항이 응시자들의 실력을 민감하게 잘 반영한다는 사실을 의미한다. 따라서 검사가 신뢰롭게 응시자들의 진점수를 반영하려면 문항의 변별력이 가능한 높아야 하고, 문항변별력이 높은 문항들로 구성된 검사는 비교적 안정적으로 개인의 특성을 측정할 수 있을 것이다.

문항분석을 위하여 자료분석의 허가를 얻은 TOEIC시험은 1999년도 2월과 3월에 실시하였던 두 번의 정기시험들이었다. 고전검사이론과 문항반응이론을 동원하여 문항들의 통계적인 특질들을 재차 조사하여 본 결과 두 검사의 평균 난이도 지수는 99년도 2월 시험의 경우 듣기시험이 .643, 읽기시험이 .626인 것으로 산출되었다. 각 문항들이 4지 선다형이었던 점을 고려하여 본다면 이들 시험의 바람직한 난이도 수준은 약 .625 정도이다. 그림 1에 제시된 2월 TOEIC 정기시험의 문항의 난이도 분석결과는 적절한 정도의 난이도 수준을 지니는 것으로 보인다. 또한 문항난이도의 분포도 약간의 부적 편포를 이루어, 사지선다형 문항의 바람직한 난이도, .625를 중심으로 어려운 문제부터 쉬운 문제들까지 다양하게 시험문제들을 출제한 것으로 나타났다.

그림 2에는 1999년도 3월 TOEIC 정기시험의 문항난이도 분석이 요약되어 있다. 난이도 분석의 결과는 2월 정기시험과 마찬가지로 상당히 바람직한 수준의 평균난이도를 지니는 것으로 나타났다. 듣

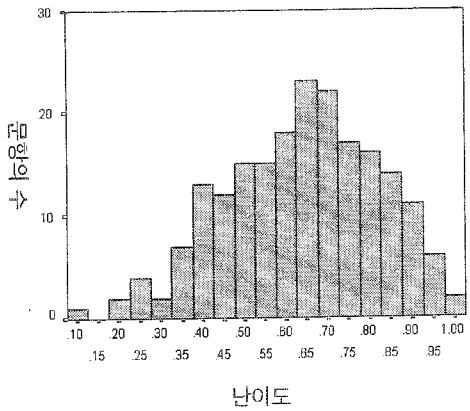

그림 1. 99년도 2월 문항난이도 분포

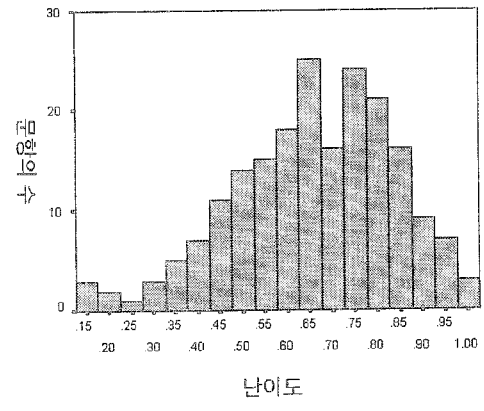

그림 2. 99년도 3월 문항난이도 분포

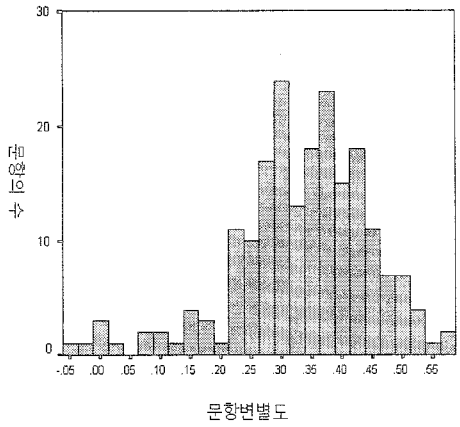

그림 3. 99년도 2월 시험 문항변별도

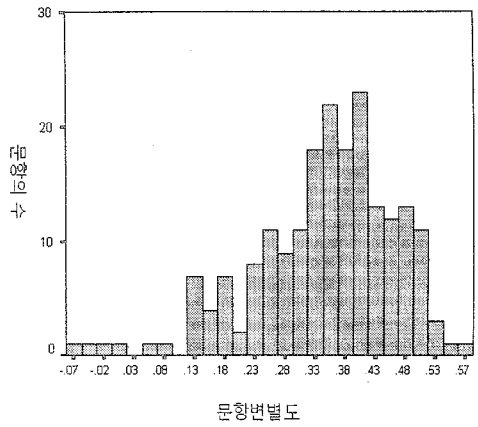

그림 4. 99년도 3월 시험 문항변별도

기시험의 경우에는 평균 난이도가 .640 읽기시험의 경우에는 평균 난이도가 .674였다. 역시 문항난이도 분포 부적으로 약간 편포되었으나 비교적 다양한 난이도를 지닌 문항들이 포함되어 있었다. 이들 난이도 분포는 현재와 같은 사지선다형 문항들의 추측 응답율을 고려하여 볼 때 적절한 형태라 결론지어졌다.

문항들의 변별도에 대한 분석으로는 SAS 프로그램으로 point biserial correlation을 산출하였다. 이 상관지수는 문항들에 대한 각 개인의 응답성 여부와 총점간의 관련성을 반영하여 준다. 만일 총점에서 영어실력이 있다고 판명 난 응시자라면 각 문항들에 대하여서도 정답을 잘 찾을 수 있어야 할 것이다. 그러나 총점 상에서 실력이 없는 응시자라 드러난다면 각 문항들에서 정답 대신 오답을 선택했어야 할 것이다.

그림 3에서 나타나듯이 99년도 2월 정기시험 중 듣기시험의 평균 변별도는 약 .472, 그리고 읽기시험에 있어서의 평균 변별도는 .460인 것으로 나타났다. 사지선다형 문항들에 대하여 바람직한 수준의 문항변별도 정도는 .2에서 .6 정도인 것을 고려하여 볼 때, 이 정도 수준의 변별도 지수는 상당히 만족스러운 것으로 여겨진다.

그림 4에는 99년도 3월 TOEIC 정기시험의 변별도 분석결과가 요약되어 있다. 듣기시험과 읽기시험 각각에 대한 평균변별도 지수는 .440, .509로 각 문항들은 개인의 영어실력을 우수하게 반영하여 주었다.

부가적으로 TOEIC 2월 시험과 3월 시험을 대상으로 문항들의 난이도와 변별도 지표들 간의 관련성을 살펴보았다. 이들 결과는 난이도와 변별도 간의 일반적인 관련성을 알려 준다. 그림들에서 발견할 수 있는 점은 문항의 난이도와 변별도는 이차함수의 관계를 지녀서, 난이도를 중간 정도로 유지시킨 경우 문항들의 변별도, 즉 실력이 있는 사람과 실력이 없는 사람들을 구분하여 주는 정도가 가장 좋다는 사실이다.

즉 너무 어려운 문항들이나 쉬운 문항들의 경우에는 오히려 문항의 변별력이 저하된다는 것이다. 따라서 어려운 문항들일수록 수험자들의 실력을 더 잘 변별할 수 있다는 통상적인 믿음은 잘못된 것이라 결론지을 수 있다. 이보다는 현재 TOEIC 시험처럼 일정하게 중간 정도의 난이도를 지니고 있는 문항들로 구성된 시험이 일반 수험자들을 대상으로는 훨씬 훌륭한 변별력을 유지할 것이라 사료된다.

고전검사이론의 방식으로 분석된 문항분석결과 외에도 문항반응이론에 따라 문항분석을 다시 한

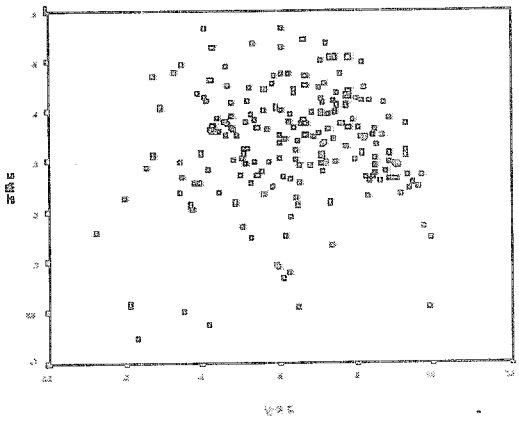

그림 5. 99년도 2월 문항난이도와 변별도

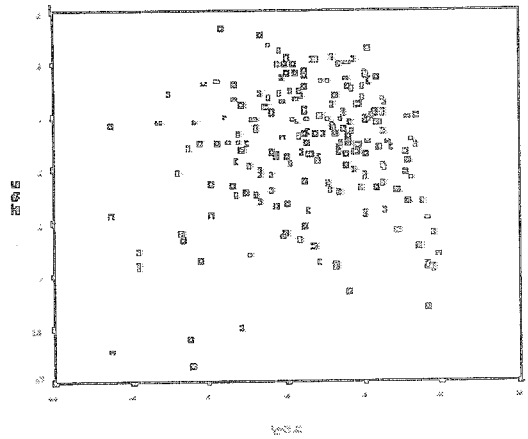

그림 6. 99년도 3월 문항난이도와 변별도

번 실시하였다. BILOG 3 프로그램을 이용하여 3 모수 모델로 분석된 문항분석 결과 역시 현재 분석에 포함되었던 두 개의 시험들은 적절한 수준의 문항특성을 지니는 것으로 확인되었다.

문항의 난이도를 의미하는 Threshold의 값과 문항의 변별도를 나타내는 Slope의 값을 산출하였다. 일반적으로 문항들이 적절한 수준의 통계적 특성을 지니려면 Threshold 값은 -3에서 3의 값을, Slope의 값은 -2.8에서 2.8의 값을 지녀야 한다(Mislevy & Bock, 1990). 현재 선정된 TOEIC 시험들의 경우, 듣기시험이나 읽기시험 모두에 있어서 대부분의 문항들이 이 범위 내의 문항변별도와 문항난이도 지수를 지니는 것으로 나타났다.

논 의

현 연구의 가장 큰 의의는 우리나라 응시자들을 대상으로 TOEIC검사의 타당도를 확인하여 보았다는 점이다. 물론 전 응시자들을 대상으로는 신뢰도나 타당도 분석이 이루어진 적이 있으나, 우리나라 응시자들만을 대상으로 문항의 특질들이 분석된 공인된 결과는 아직까지 보고된 적이 없다(International Communication, 1998). 최근 응시자들의 폭발적인 증가로 TOEIC 검사의 일반화 추세는 그 검사 자체에 대한 여러 가지 질적 의구심을 자아내었었다. 현 연구는 이와 같은 일상적인 의문에 경험적인 증거를 제시하여 준다.

결과를 요약하자면, 현재 분석에 포함되었던 TOEIC 검사들은 비교적 적절한 수준의 타당도, 신뢰도 및 문항적 특질을 지닌다고 결론지을 수 있다. 그러나 점수의 안정성에 있어서는 아직도 개선되어야 할 측면이 있다. 평균적으로 볼 때 총점에 있어서 약 18점 정도의 증가 추세는 일반적인 세인들의 기대보다는 그리 높지는 않다. 그러나 그 원인에 대하여서는 문제의 출제자 측에서 좀더 신중한 고려가 있어야 할 듯하다.

현 분석에 포함되었던 자료들은 국내에서 추적이 가능한 것으로는 비교적 방대한 양이었다. 그러나 신뢰도 지수들이나 타당도 지수들을 산출함에 있어서 각기 포함된 자료들이 서로 달라 현재의 상관계수들을 확고부동한 지표들로 취급하기에는 상당한 무리가 있다. 이들 분석들은 사전에 엄격한 통제하에 디자인된 것이기보다는, 각기 적절한 지표를 산출할 근거에 맞추어 사후에 수집된 것이다. 따라서 개별 분석에 포함되었던 검사들이 서로 다른 경우도 있었다. TOEIC검사는 기본적으로 equipercentile equating 방식으로 채점된다. 따라서 동일한 백분위

의 응시자들의 점수는 동일한 양의 진점수를 반영하고 있다고 가정되었다. 현 연구에서는 이런 가정을 토대로 다양한 분석을 시도하였으나, 추후에는 이 가정에 대하여서도 좀더 주의깊은 확인 작업이 동반되어야 할 것이다. 미래의 연구들에서는 검사의 동등화 이론이나 일반화 가능도 이론을 적용하여 개별 검사들 간에 공유된 진분산을 정확하게 추정한 상태에서 이와 같은 분석을 수행하여야 할 것이다.

새롭게 다가오는 천년에 대해 많은 전문가들은 지식기반 사업이 가장 훌륭한 전망을 지녔다고 예견하였다. 이렇게 보자면 인간의 개인차를 측정하는 여러 가지 검사들의 개발은 지식기반 사업 중 가장 대표적인 예가 될 수 있을 것이다. 개인차를 측정하는 물리적인 지표들의 개발과 과학기술의 발전은 인간의 특성에 대한 진단, 선발, 배치 등 모든 인사에 중요한 경험적 지표가 되는 검사의 형태를 지금까지의 지필검사로부터 수행검사의 양식으로 바꾸어 가고 있다. 따라서 이들 검사분야의 발전에도 도약의 시기가 인접했다고 예측할 수 있을 것이다. 그러나 진보된 형태의 검사양식의 도래에도 불구하고 한가지 꼭 유의하여야 할 점은 검사의 '질'에 대한 의문이다. 검사의 양식이 무엇이건 앞으로도 여전히 인사의 선발이나 배치 및 진단의 목적으로 검사들을 사용하는 한, 그것의 '질'에 대한 의문은 계속적으로 제기되고 도전될 것이다. 검사의 판별력이나 예측력에 대한 이런 질의는 검사양식의 세련화와는 별개의 문제이다.

TOEIC 검사의 경우가 바로 세인들의 이런 종류의 질의에 대한 좋은 사례가 되어 왔다. 최근 TEPS라는 경쟁 시험의 등장은 TOEIC 검사의 통계적 자질에 대한 도전장을 던졌다. 현 연구는 이런 종류의 의구심을 어떤 방식으로 해결해 나가야 할 지에 대한 혜안 한 가지를 제공하여 주고 있다. 많은 검사관련자들은 이 자료를 TOEIC이라는 시험의 타당도 지표로 유용하게 이용할 수 있을 것이다. 그러나 과연 TOEIC이 우리나라 표본에 진정으로 적합한 시험인가에 대하여서는 수많은 이후의 타당도 연구들이 수행되어야만 대답될 것이다. 타당도와 신뢰도에 관한 반복적인 검증, 표준화나 검사 동등화 과정에 대한 심층적인 분석들이 계속 이어져야 할 것인 바, 이유는 각각의 분석이 그때마다 포함된 표본이나 상황의 특수성을 그대로 반영하기 때문이다. 이 같은 문제는 꼭 TOEIC과 같은 대형검사가 아니더라도 거의 모든 심리검사에 대해 제기될 의문인 바, 앞으로는 검사의 개발 및 적용 이외에 검사의 경험적 특질에 대한 보완 자료들의 수집이 동반되어야 만이 특정한 검사들이 생존해 나갈 수 있게 될 것이다.

참 고 문 헌

국제교류진흥회 (1995). Guide for SEPT Users. 서울: 국제교류진흥회.

국제교류진흥회 (1996). Guide for TWT Users. 서울: 국제교류진흥회.

Anastasi, A. (1990). *Psychological testing.* New York: Macmillan.

International Communication Foundation (1998). *News letter*, 9 (3), 1-15.

Mislevy, R. J. & Bock, R. D. (1990). *Bilog 3: Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software Inc.

Woodford, P. E. (1982). The test of English for international communication (TOEIC). In C. Brumfit (Ed.), *English for international communication.* New York: Pergamon Press.

International Communication Foundation (1998). *News letter*, 9 (3), 1-15.

韓國心理學會誌: 社會問題
Korean Journal of Psychology: Social Issues
1999, Vol. 5, No. 2, 49~58

A Validation Study on TOEIC

Lee, Soo Jung

Kyonggi University, Division of General Studies

This study investigated various statistical characteristics of TOEIC only based on Korean samples, which have rarely been studied before. Test-retest reliability indices produced with 3-month duration and KR20 based on samples included suggested both scales of TOEIC having good quality of score stability and internal consistency. The excellence of criterion-related validity was found based on the relationship with other tests measuring similar domains and a test measuring practical English use in self-report survey form. The item analyses also showed TOEIC tests included in this study had a satisfying level of item difficulty and discrimination.