Computerized Adaptive Language Testing 알고리즘 개발 및 타당도 검증

최 인 철
(고려대학교)

Choi, Inn-Chull. (2004). Development and validation of a computer adaptive language testing algorithm. *Language Research* 40(1), 227-252.

In the era of information and communication technology, where computerized adaptive language testing continues to draw a keen attention from language testers, the present research attempts to develop a reliable and efficient algorithm of computerized adaptive language testing (CALT) and to investigate the extent to which the algorithm results in reliable and valid outcome. Employing an eclectic approach to analyzing real data as well as simulation data, the research has developed a highly reliable and valid algorithm on the basis of fundamental considerations including the difficulty level of initial items, the estimation method of the ability parameter, the item selection, the test length, the measurement error, the context effect, the item-based response time, etc. Focusing on the concurrent validity, the overall research findings suggest that the CALT version of the TEPS (Test of English Proficiency developed by SNU) accommodating the item-based response time factor is superior to the paper-based language test (PBLT) version of the TEPS in terms of measurement accuracy of overall English proficiency. Furthermore, the correlational analyses based on test-takers' performance of CALT, PBLT, and TOP (the Test of Oral Proficiency) strongly support the plausibility of the test methods of the TEPS in which the grammar test is administered in a speeded manner and the listening comprehension test input is presented in a purely oral mode.

Key words: 컴퓨터 개별 적응 언어능력 시험, 능력모수 추정, 문항 추출, 측정오차, 타당도, 신뢰도, Computerized Adaptive Language Test: CALT, estimation of the ability parameter, item selection, measurement error, validity, reliability

1. 서 론

컴퓨터로 수험자 개인별로 수준에 알맞은 시험을 치름으로써 타당한 수험결과를 얻고자 개발된 개별 적응 검사 (Computerized Adaptive Testing: CAT) 분야에서는 어느 정도 많은 연구가 이루어졌으나 (백순근 외, 1998; Mazzeo & Harvey,

1988; Mead & Drasgow, 1993; Russel & Haney, 1997; Sands, Waters, & McBride, 1997; Wainer, et al., 1990; Wainer, et al., 2000), 언어 테스팅 분야에서 특히 수험자 실력수준과 시험 방식 및 제시방법의 상호작용성이 매우 큰 컴퓨터 개별 적응 언어테스팅 (Computerized Adaptive Language Testing: CALT) 분야에서는 체계적인 연구가 매우 부족한 상태이다 (Brown, 1997; Dunkel, 1991). 지필 고사 방식과 많은 차이를 보이는 멀티미디어적으로 시험 내용을 제시하는 시험 방식에 대한 이해가 부족하므로, 특히, 시험 방식 양상 (Test Method Facet: TMF)이 매우 중요한 역할을 하는 언어테스팅 분야에서는 CALT의 시험 방식 개발 연구와 지필고사-CALT간의 상호 비교성 연구 (Choi, Kim, & Boo, 2003)등 체계적인 타당성 검증 연구가 요구된다.

이런 점에서, 본 연구에서는 현재의 CAT의 현황과 평가 이론적으로 핵심적인 고려사항 (최인철, 2000; Brown, 1997)에 근거해서, 타당도 높은 시험 방식과 신뢰도 높은 수험자 능력 측정 결과를 얻을 수 있는 IRT (Item Response Theory)에 근거한 평가 알고리즘의 이론적 모델을 시뮬레이션 방법과 실제 피험자 자료를 토대로 CALT의 이론적 모델을 개발하였다. 여러 고려사항, 즉, 1) 시험시작 방식, 2) 문항 난이도와 변별력 고려한 최적의 문항 추출 알고리즘, 3) 합리적인 능력 모수 추정 방법, 4) 문제은행 및 IRT 모델 선정, 5) 내용/소재 영역, 유형 및 난이도 등을 고려한 내용 균형화 (content balancing), 6) 시험 정보함수, 표준측정 오차, 시험 길이 및 수험 시간을 종합한 시험 종료 방식, 7) 총 시험 시간 및 문항 응답 시간 등을 고려한 수험 결과 조정 등등을 고려하여 신뢰성있고 타당한 CALT 알고리즘을 개발하는 연구를 수행하였다.

2. CALT 알고리즘 개발

CALT의 타당한 알고리즘의 이론적 모델을 개발하기 위한 다음과 같은 여러 가지 변수를 고려하고 정규 분포의 능력을 지닌 1000명의 가상 수험자를 대상으로 무작위 표본 추출하는 시뮬레이션 알고리즘을 통한 분석 연구 (최인철, 2000)를 수행하였다. 또한, 이 시뮬레이션 결과를 바탕으로 개발된 CALT 알고리즘에 TEPS (Test of English Proficiency developed by Seoul National Univ.) 시험 문제은행을 접목한 전산 프로그램을 평가 도구로 하여 서울시내 소재 대학의 영문과 및 대학영어 수강생 49명을 대상으로 보다 체계적인 실험 연구를 2001년에 실시하여 타당한 알고리즘을 2002년에 개발하였고, SOPI (Simulation Oral Proficiency Interview) 방식의 구술시험인 TOP (Test of Oral Proficiency: 최인철, 2000; Choi, 1998) 시험을 통해 CALT 알고리즘의 타당성 검증 연구를 수행 하였다.

본 연구를 위해서 시행된 알고리즘 개발의 수행 방법과 결과에 대해서 구체적으로 1) 초기 문항 추출방법 (문항수와 난이도), 2) IRT 모델 선정, 3) 능력 모수 추정 방법 (MLE (Maximum Likelihood Estimation: 최대우도추정법), Bayesian (베

이지언 방법)), 4) 문항 추출 방법 (기계적 분지 모델, 다단계 고정모형 (Multistage Fixed-branching CAT Model), 수학적 최적화 모델, 절충적 모델, 실제 수험자 응답 결과), 5) 시험 길이, 6) 종료 방법 (측정오차, 문항수, 수험/응답 시간, 혼용 방법) 등의 순서대로 제시한다. 마지막으로 TOP의 수험결과와 지필고사 수험 결과를 바탕으로 CALT 타당도 검증의 연구 결과를 제시하였다. CAT의 현황과 CAT의 장점과 제약점, 그리고, CALT의 중요 고려사항에 대한 자세한 내용은 Brown (1997)과 최인철 (2000)을 참조하기 바란다.

2.1. 초기 문항

2.1.1. 문항수와 난이도

능력 모수를 추정하기 위하여 EM (Expectation Maximization) 루프를 돌리지 않고, 수험자의 능력초기 값을 결정하기 위하여 어느 정도의 난이도 수준으로 몇 개의 문항을 제시하는 것이 가장 합리적인가를 분석하였다. 우선, 25문항의 시험 문제를 가정하여, 초기 문항을 4, 5, 6, 7, 8개로 5가지 변수로 제한하였고, 처음 추출될 때의 문항의 난이도 b값의 범위를 -.5~+.5, 0~-1, -.5~-1.5, 1~-1의 네가지 영역으로 제한하고, 수험자가 문항을 맞출 확률의 범위를 .4~.6로 정하고 시뮬레이션을 시행했다. 모든 경우의 수에 관한 표는 일관성있는 결과를 보이므로, <그림 1> - <그림 6>은 최초 문항 4, 5, 6의 3가지 경우와, 난이도 b 값의 범위 -.5~+.5, 0~-1의 2가지 경우가 혼합된 6가지의 경우만 제시한다.

결과 분석의 주요 지표로서는, 실제 수치와 모델에서 추정하는 이론적인 수치의 일치도를 의미하는 지수인 RMSE (Residual Mean Square Error)와, 시뮬레이션의 수험자 실력과 최종 추정된 수험 결과간의 상관관계 계수를 활용하였다. 상관관계 계수는 매우 높은 수치를 보이고 있는데, 알고리즘이 잘못되었을 때에는 상관관계 계수가 매우 낮게 나오므로, 이런 상관관계와 RMSE의 분석은 의미있는 연구이다. 아래의 <그림 1> - <그림 6>를 보면, 초기 문항수가 4, 5, 6개의 경우 큰 차이가 없었지만, 초기 문항수가 적을수록 (즉, 수험자 실력에 적응하는 문항이 많을수록) RMSE가 근소한 차이로 적게 나타났고, 상관관계가 매우 근소한 차이로 좋게 나타났다. 또한, 초기 문항의 난이도 범위는 ∴5~+.5 사이가 좀더 좋은 결과를 나타낸다. 따라서, 초기 문항의 수는 4개로 하고, 난이도 범위는 ∴5~+.5로 정하여, 실제 수험자 시험 자료를 얻기 위한 알고리즘을 개발하였다.

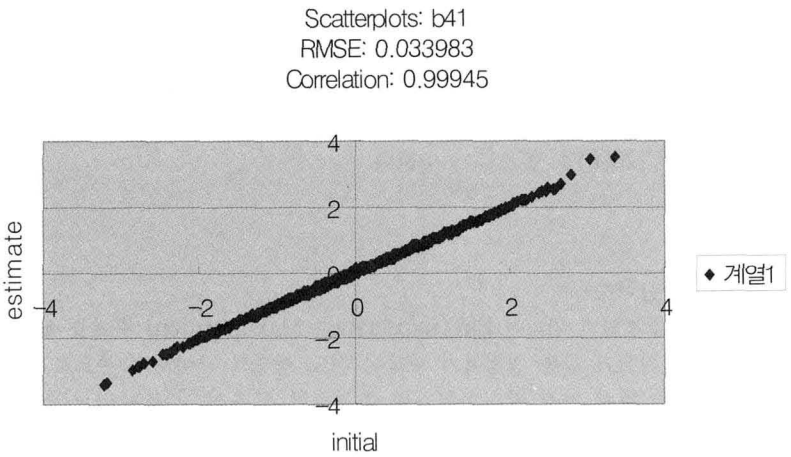

그림 1. 시뮬레이션 결과 <초기 문항수 4; b: 0~-1>

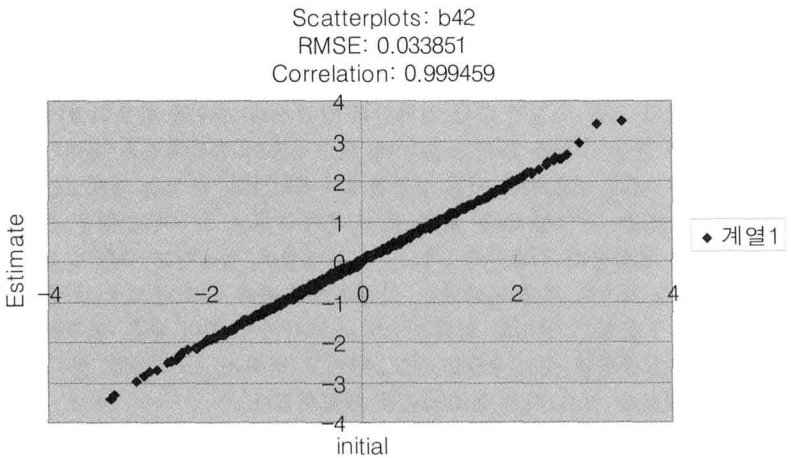

그림 2. 시뮬레이션 결과 <초기 문항수 4; b: -.5~+.5 >

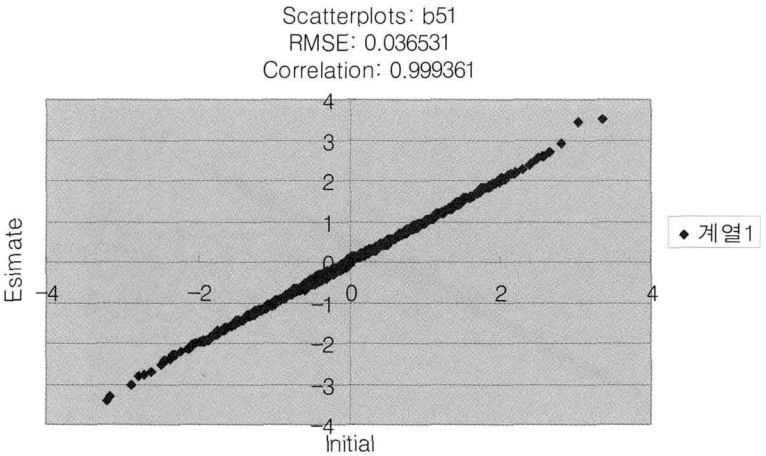

그림 3. 시뮬레이션 결과 <초기 문항수 5; b: 0~-1 >

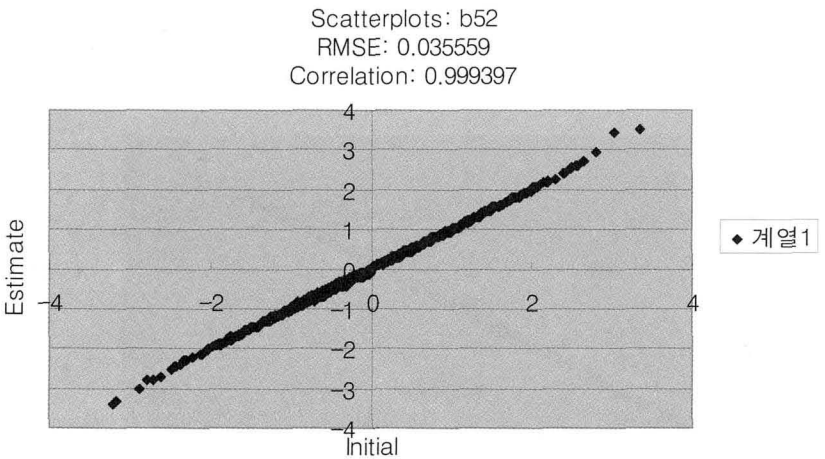

그림 4. 시뮬레이션 결과 <초기 문항수 5; b: -.5~+.5>

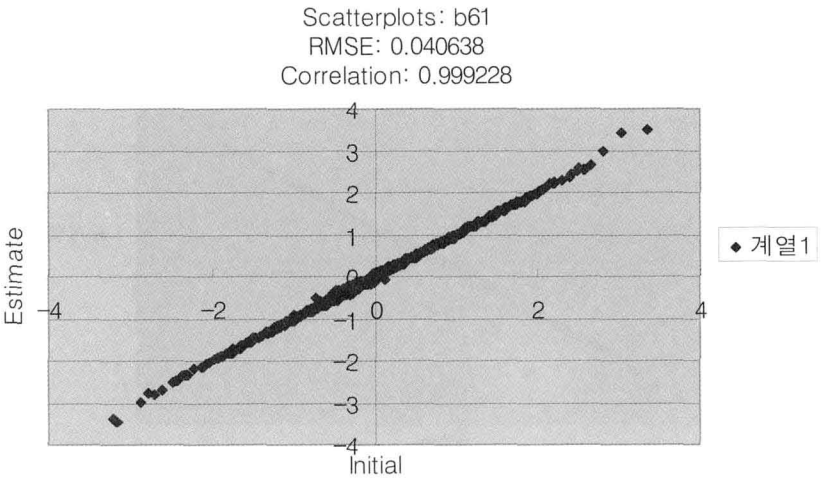

그림 5. 시뮬레이션 결과 <초기 문항수 6; b: 0~-1 >

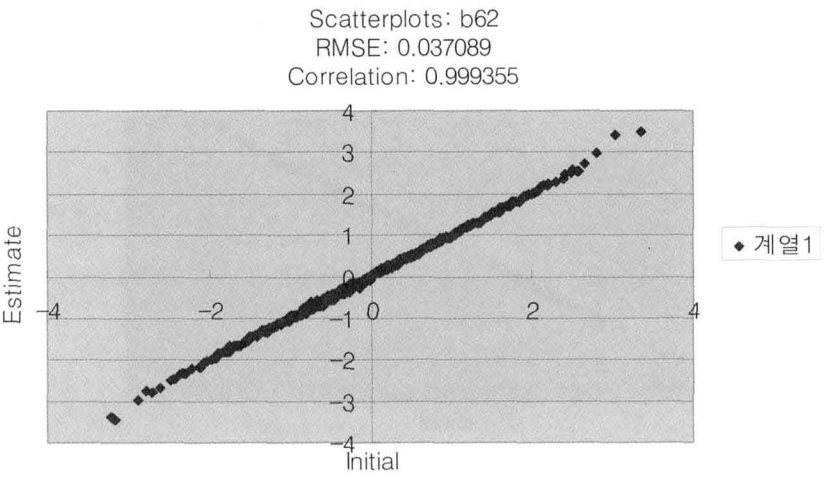

그림 6. 시뮬레이션 결과 <초기 문항수 6; b: -.5~+.5 >

2.1.2. 보안문제

다음과 같이 문항수의 변화를 통해서 보안문제를 해결하는 것이 바람직하며, 두 가지 방법상에서 신뢰도면에 큰 차이는 없는 것으로 나타나고 있다 (Wetzel & McBride, 1986). 1) 고정 문항수 (fixed set size) 방법은 가장 타당한 문항 수를 고정해서 선택한 후, 한 문항씩 임의로 추출하는 방법으로서, 문항 수가 많아지면, 정확도가 떨어지는 문제점이 있다. 2) 감소 문항수 (shrinking set size) 방법은 문항이 매번 추출될 때마다 전체 문항수가 순차적으로 감소하는 방법으로서, 감소 정도를 변화시켜서, 임의추출 정도에 변화를 주는 방법이다. 본 연구에서는 신뢰성높은 알고리즘 개발자체에 초점을 두었기 때문에 보안문제 자체에 대해서는 고려하지 않았다.

2.2. 문제은행과 IRT 모델

본 연구에 활용된 TEPS 문제은행의 모든 문항은 IRT의 적용 분석을 위한 전제조건인 일차원성 (Unidimensionality)이 Stout's Dim test (Stout et al., 1991)을 통하여 점검되었고, 일차원성 전제조건을 만족되는 문항들로 구성되어 있다 (Choi, 1994; Choi, 1995). 또한, TEPS의 문제형식은 하나의 독해/청해 지문, 또는 문법/어휘 문두 (stem)에 하나의 문항이 제시되는 일지문 일문항 (One Passage One Item) 원칙을 지킴으로써 IRT의 대전제인 국부 독립성 (Local Independence)을 내재적으로 만족하도록 되어있다 (Choi, Kim, & Boo, 2003).

IRT 모델중에서 1 PL (Parameter Logistic)/ Rasch 모델은 변별도를 고려하지 않고 난이도만을 근거로 채점하기 때문에 채점의 측정오차가 큰 것으로 알려져 있으므로 (Choi & Bachman, 1992; Hambleton & Swaminathan, 1984; Hulin, Drasgow, & Parsons, 1983), 본 연구에서는 고려하지 않았다. 또한, CAT에서는 원칙적으로 수험자의 능력수준에 적절한 곤란도의 문항을 제공하므로 추측도의 영향이 미미할 것이므로, 수험자의 개인 실력 수준에 무관하게 모든 문항 추측도를 동일하게 가정하는 3 PL 모델은 수험자의 실력 수준에 적응시키는 시험인 CAT에서는 바람직하지 않다고 판단한다.

지필고사의 경우에서도 대규모로 실시된 언어테스팅 연구의 자료 분석 결과에 의하면, 2 PL 모델이 3 PL보다도 모델-자료 일치도면에서 거의 차이가 나지 않는 경우가 많았으며 (Bachman et al. 1995; Choi, 1992), CAT경우에서는 더욱더 그런 현상이 나타날 것으로 판단된다. 따라서, 본 연구에서는 2 PL 모델을 활용한 문항 속성 추정치를 활용하여 구축한 TEPS 문제은행 DB를 활용하였다.

2.3. 능력 모수 추정 방법

수험자 능력 추정 방법에는 크게 다음과 같은 두가지 방법이 있는데, 모두 장·단점을 가지고 있으므로, 개발하는 시험의 목적과 방식을 고려해서 가장 적절한 방법을 선택해야 한다 (Hambleton, Swaminathan, & Rogers, 1991).

2.3.1. MLE (최대우도추정법)

피험자의 능력 수준을 추정하는 시점에서, 최대의 문항 정보를 제공할 수 있다고 판단되는 문항을 선택하는 방법으로서, 모두 정답을 한 만점의 경우와 모두 오답을 한 영점의 경우에 확률적으로 계산 불능의 문제가 있다. 이런 단점의 해결책으로서는 (1) 수험자의 응답중 적어도 정답과 오답이 하나 이상 나올 때까지 Bayesian 추정법을 사용하다가 MLE으로 전환하거나, (2) 수험자의 응답중 적어도 정답과 오답이 하나 이상 나올 때까지 제시하는 문항의 곤란도 수준을 일정한 크기 (예: logit 0.3)로 증가시키거나, 감소시키거나, (3) 시험 완료후 만점자에게는 매우 높은 능력 추정치를 (예: logit 5.0), 영점자에게는 매우 낮은 능력 추정치 (예: logit -5.0)를 임의로 부여하는 방법이 있을 수 있다 (Weiss & Kingsbury, 1984; Weiss, 1983).

본 연구에서는 처음 4개 문항을 다 맞춘다면, 초기 모수치를 .5로 설정하고, 틀리는 문항이 나올 때까지, 문제은행의 문항 난이도 간격 (예: .125)씩 부여하였으며, 반대의 경우는 초기 모수치를 -.5로 설정하고, 동일한 간격으로 감점하는 알고리즘을 활용하였다.

2.3.2. Bayesian (베이지언 방법)

수험자 능력의 사전 정보를 이용하여 수험자의 능력 모수치 추정에 있어서, Bayesian 사후 변량 (Baysesian posterior variance)을 최소할 것으로 예측되는 문항을 선택하는 방법으로서, 한 문항에 대한 수험자의 응답을 근거로 Bayesian 추정 방법에 의해 수험자의 능력 수준이 재추정되고, 그 시점에서 다시 아직 시행되지 않은 문항중 사후 변량을 최소화 할 문항을 찾게 된다. 사전에 정해둔 사후변량의 수준에 도달할 때까지 반복하게 된다 (Owen, 1975). 이때에 능력추정치의 사후확률분포는 정상분포를 가정한다. 이론적인 장점으로는, 수험자의 사전 정보를 이용하여 보다 수험자 능력의 보다 정확한 추정이 가능하다. 그러나, 이론적으로 볼 때에 회귀 (regression) 효과가 발생하여, 추정치가 가운데로 축소되는 경향이 중요한 단점으로 지적되고 있다. 또한, 현실적인 단점으로서는 수험자의 능력의 초기 정보 획득이 실제로 거의 불가할 뿐만 아니라, 최초/ 사전 추정치의 정확성 여부에 따라서 최종 추정치가 편파적일 가능성이 높다 (Sands, Waters, & McBride, 1997).

2.4. 문항 추출 방법

어떤 방법으로 문항을 추출할 것인지 바람직한가에 대한 여러 선행 연구를 비교 분석하여, 가장 타당한 알고리즘을 결정하였다.

2.4.1 기계적 분지 모델

1970년도까지 많이 사용했던 문항 추출 방법으로서, 융통성이 부족한 것이 심각한 문제점으로 나타나며, 1) 유동적 수준 검사 (flexilevel test: Lord, 1971a), 2단계식 검사 (two-stage test: Lord, 1971b), 2) 피라밋식 검사 (pyramidal test: Hansen,

1969), 3) 단계적 개별적응 검사 (stradaptive test: Weiss, 1974)등의 기법이 있다.

2.4.2. 다단계 고정모형 (Multistage Fixed-branching CAT Model)

기계적인 분지 모델의 문제점을 보완하여 다단계 고정 모형이 개발되었다 (Wainer & Kiely, 1987). 개별문항을 하나의 문항군 (testlet: 단일 내용 영역)으로 대체하고, 문항군은 모든 수험자에게 동일 문항들이 제시되는 선형 (linear)과 수험자의 사전 응답에 따라서 각 수험자에게 다른 문항이 제시되는 위계형 (hierarchical)으로 구분한다 (Kingsbury & Zara, 1989; Thissen et al, 1989).

2.4.3. 수학적 최적화 모델

매 문항마다 수험자의 능력 수준이 결정되면, 그 능력수준에 근거해서 문항 정보가 가장 큰 다음 문항이 추출되는 수학적/통계적 모델이다. 여기에는, Owen (1975)의 Bayesian 채점 방식과, MLE (최우도추정방법)에 근거한 Lord (1977)의 BRTT (Broad Range Tailored Test) 방법 등이 있다.

2.4.4. 절충적 모델

본 연구에서는 문제 은행의 문항수가 부족한 여건이므로, 순수하게 하나의 모델, 즉, 수학적 최적화 모델로 추정하는 것보다 기타 방법과 함께 절충적인 접근법을 취하는 것이 바람직하다. 따라서, 최초 문항을 추출할 때에는 기계적 분지 모델과 문항 정보의 극대화하려는 개념에 근거한 수학적 최적화 모델의 적절한 혼합을 꾀하였다. 최종 시뮬레이션 결과에 의하면, 가상적인 실제 점수와 알고리즘에 근거한 추정 점수간의 상관 관계는 그림 7에서 보이는 바와 같이 .998786으로서 매우 높은 것을 알 수 있다.

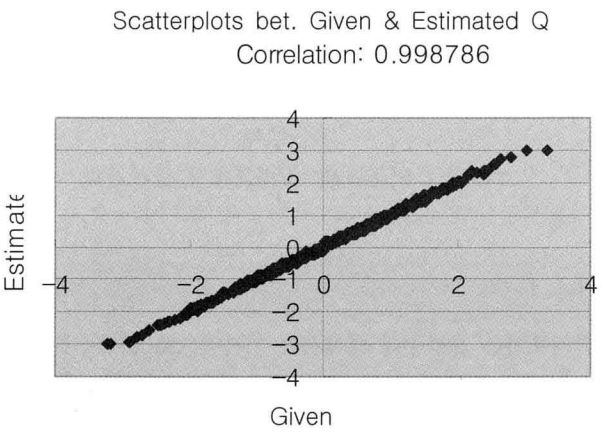

그림 7. 시뮬레이션 결과

2.4.5. 실제 수험자 응답 결과

TEPS성적이 4급부터 1+급까지 다양한 실력 수준을 정규분포로 보이는 실제 수험자 49명이 본 연구에서 개발한 알고리즘으로 운영되는 TEPS의 CALT를 치루어서 나온 수험 결과와 TEPS의 PBLT (Paper-based Language Test)의 수험 결과의 전체 점수와, 각 영역별 상관 관계는 표 1과 같다. 각 영역별 상관 관계 계수도 독해를 제외하고는, 모두 .8이상으로서 높은 수준을 보이고, 전체점수간의 상관관계는 .904로서 상당히 높은 것을 알 수 있다. (실제 수험자의 반응과 채점 결과를 기록한 내용은 "부록 CALT 알고리즘 능력 모수 추정 결과 예" 참조바람.)

표 1. PBLT 총점-CALT 영역점수 상관관계

	총점	청해	문법	어휘	독해
PBLT	.90	.82	.83	.90	.75

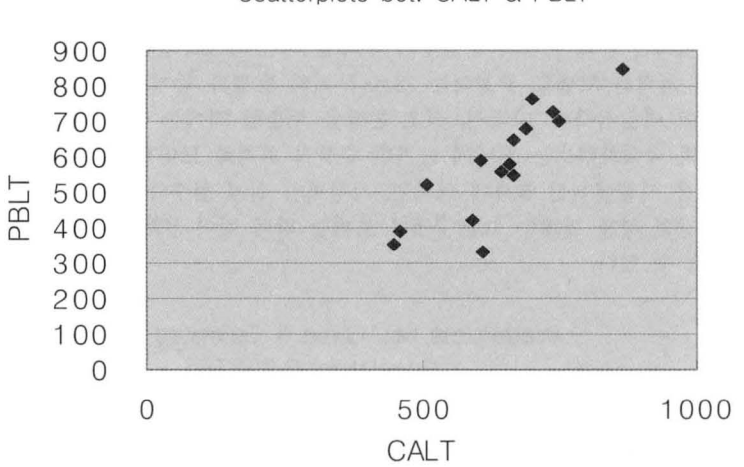

그림 8. CALT총점 - PBLT총점 상관관계

2.5. 시험 길이

시험 길이를 좌우하는 문항수에 대해서는 기본적으로 문항수를 고정시키는 경우와 변화시키는 두가지 방법이 있을 수 있는데, 일반적으로는 혼합 방식을 취하는 경향이 많은 것으로 알려져 있다. 수험자마다 상이한 문항수의 시험을 보게 된 후에, 다른 결과를 얻게 된 경우에 자신의 실력 때문이 아니라, 문항수의 차이에서 신뢰도에 많은 영향을 미치게 되어서 수험 결과가 다르게 나왔다고 주장할 때에

시험 시행자 입장에서는 시험 길이 방식에 대하여 논리적으로 완벽하게 변호하기 어렵게 된다.

따라서, 배치고사 (placement test)같이 대 사회적인 책임이 비교적 적은 시험에서는 시험 길이를 CAT의 이론적인 측면만을 고려한 가변적인 시험 길이 방식을 채택하는 경우가 있으나, 대사회적인 책임이 큰 시험에서는 일반적으로 시험의 길이를 고정하는 것이 바람직한 방법으로 판단된다. 이론적으로는, 몇 문항의 시험 길이가 가장 합리적인 지에 대해서는 아래에서 기술하는 다양한 종료 방법과 시험 영역/소재별 내용, 시험유형, 난이도 등을 종합적으로 고려한 내용 균형화 (Content Balancing) 연구에 근거하여 결정하는 것이 바람직할 것이다 (Wainer & Kiely, 1987). 이런 점등을 고려하여 본 연구에서도 시험의 길이를 고정하는 것으로 설정하였다. 최인철 (2000)에 의하면, 현재의 지필고사 시험 길이 (200문항)를 반 (100문항)으로 줄여도 안정적인 신뢰성이 나오므로, 본 연구에서는 지필고사 문항수의 반 (100문항: 청해 30문항, 문법 25문항, 어휘 25문항, 독해 20문항)으로 길이를 고정하여 시험을 제작하여 연구의 평가도구로 활용하였다.

2.6. 종료 방법

시험의 길이를 결정하는 시험 종료 알고리즘을 개발할 때에는 내용 균형화와 함께 다음과 같은 변수 -- 측정오차, 문항수, 응답시간 등을 고려한다.

2.6.1. 측정오차

측정 오차의 수준을 고정시키는 방법으로서, 예컨대 95% 신뢰구간이 logit scale +-0.3 이내로 적어질 때 종료할 수 있으며, 어느 정도 (예: 20여개)의 충분한 문항수를 제시하면, 시험의 신뢰도에 특별한 문제가 없는 이상 문항수가 증가함에 따라서 측정 오차는 점차 줄어들며, 보통 .1정도로 수렴한다. 측정오차의 역으로서 계산되는 시험 정보 (test information)를 기준으로 시험 종료 알고리즘을 개발할 수도 있다. 본 연구에서도 측정 오차가 .1이하가 되지 않으면, 시험정보가 미흡하여 신뢰성 있는 능력 모수를 추정하지 못하는 것으로 간주하여, 그런 수험자의 경우에는 측정대상 능력에 대한 평가결과를 신뢰할 수 없다는 경고를 제시하였다.

2.6.2. 문항수

최대 검사 문항수를 고정시키는 방법으로서, 시험에서 다룰 영역, 소재, 유형, 난이도별 문항의 수를 고려해서 결정하는 것이 바람직하다. 최소 몇 문항 이상을 풀게 하되, 최대한으로 풀 수 있는 문항수를 고정시키는 방법도 있을 수 있으나, 그 바람직한 범위를 결정하는 데에는 많은 시행착오가 필요하다. 시험 길이에서 언급한 바대로, 대사회적 책임이 큰 시험에서는, 수험자 개인별 수험 결과의 신뢰성, 객관성, 형평성등을 유지하기 위해서 문항수를 동일하게 제시하는 것이 바람직하다.

2.6.3. 수험/응답 시간

수험시간을 고정시키는 방법으로서, 보통 문항수를 고정시키는 동시에 최대 허용 시간도 고정시키는 방법을 활용한다. 살아 있는 의사소통능력을 타당도 높게 측정하기 위해서는, 속도화시험 (최인철, 1997; Oller, 1995)의 핵심 요소인 수험 시간을 가장 중요한 변수로 고려하는 것이 매우 중요하다. 본 연구는 문항별로 응답 시간을 기록하여 수험 결과에 반영하는 알고리즘에 근거하므로, 문항별 응답시간은 무제한 주되, 현실적인 시험 시행의 제한점을 고려하여, 전체 영역별 시험에서는 수험 시간을 제한하여서 실험을 시행하였다.

문항별 응답 시간을 고려하여 능력 모수 추정치를 문항마다 달리 계산하여서 얻어진 결과를 보면, 응답시간의 고려가 타당하다는 점을 잘 알 수 있는 특기할 사항이 관찰되었다. 즉, 미국에서 초중학교 과정에서 7년간을 교육받은 유창한 영어구사력을 소유한 피실험자가 TEPS 지필고사의 성적이 910점 (1+급)을 얻었는데, 본 연구의 CALT 시험에서는 1000점 만점을 얻게 되었다. 즉, CALT 시험에서 시간변수를 고려하여 점수를 조정한 알고리즘에 의해서 채점이 실시되므로 지필고사에서 보다 더 높은 점수를 얻게 되었다. 이는 원어민에 가까운 의사소통능력을 구사하는 피실험자가 표현 능력이 직접적으로 측정되지 않는 지필고사에서 보다, 시간변수를 고려함으로써 살아있는 의사소통 능력을 좀더 잘 측정하는 CALT에서 더 높은 점수를 얻게 된 것으로 해석된다. 이런 결과는 문항별 응답시간 변수를 고려하는 CALT의 알고리즘이 타당함을 강하게 시사하는 것으로 풀이된다. CALT 알고리즘의 타당성에 대해서는 본 연구의 III. CALT 타당도 검증에서 좀더 자세히 다룬다.

2.6.4. 혼용 방법

앞의 여러 가지 변수를 혼합 절충하여 시험의 목적과 상황에 적합하도록 개발하는 방법도 있을 수 있다 (Thissen & Mislevy, 1990).

3. CALT 타당도 검증

3.1. 전반적 의사소통능력과의 상관관계

수험자간의 능력 차이를 얼마나 타당하게 변별하는지, CALT의 수험 결과와 실제 의사소통 상황에서 발현되는 실제 능력과 비교하는 사후 검증이 필요하다. 실제 상황에서 정확하게 측정된 수험자의 의사소통 능력과 CAT의 수험 결과의 일치도에 관한 분석은 언어테스팅분야의 가장 중요한 CALT의 타당도 검증연구이다. 본 연구를 위해 사용된 CBLT 시험의 내용 타당도와 PBLT와 CBLT의 호환가능도 (comparability)는 상당히 높은 것으로 밝혀졌으며, 이에 관련된 타당성에 관한 체계적인 연구도 이미 이루어졌다 (Choi, Kim, & Boo, 2003).

이를 위해서 전반적인 의사소통 능력을 측정하는 평가 도구로서 일반화가능도 (generalizability; Brennan, 1992) 이론과 다차원적 요소의 오류분산을 점검하는 FACETS (Linacre, 1994, 1999) 분석에 근거하여 측정대상 언어요소와 과제 및 채점의 신뢰성이 높은 것으로 입증된 TOP (Test of Oral Proficiency)를 활용하였다. 제약이 많은 상황에서 CALT 알고리즘의 타당성 연구에 참여한 수험자중 자원하는 수험자 35명의 의사소통능력을 TOP의 채점 교육을 받은 2명의 원어민 (inter-rater reliability index: .9425)이 측정한 결과와 본 연구의 CALT 수험결과의 상관 관계를 조사하였다. TOP와 CALT, 그리고 TOP와 PBLT와 상관관계를 비교한다면, CALT와 PBLT 시험방식간의 전반적인 의사소통능력에 대한 평가의 타당도를 알 수 있다. 또한, 시간변수, 즉, 문항별 응답 시간을 고려한 CALT 채점 방식과 고려하지 않은 채점 방식간의 차이를 비교해 봄으로써, 어느 방식이 전반적인 의사소통능력을 더 타당하게 측정하는지도 알 수 있다.

아래의 상관관계 결과에서 볼 수 있는 바대로, 결론적으로 CALT는 지필고사보다 전반적인 의사소통능력을 더 정확하게 측정하고, 문항별 응답시간을 고려한 CALT 채점 방식이 더 타당함을 알 수 있었다. 여러 변인들의 상관관계를 좀더 구체적으로 살펴보면 다음과 같다.

CALT와 PBLT간의 타당도를 비교해 보면 다음과 같다. 모든 영역에서, CALT와 전반적 의사소통능력 (TOP의 5가지 능력 요소 점수) 상관관계 (시간고려: .9160; 시간 비고려: .9100)가 PBLT와 TOP 상관관계 (.8192) 결과보다 높게 나왔다. 비록 다소 적은 수의 피실험 집단이기는 하지만, 이로써 본 연구에서 개발된 컴퓨터 개별 적응 시험의 알고리즘이 지필시험 방식보다 전반적인 의사소통능력을 좀더 타당하게 측정할 수 있는 가능성을 보여주는 결과로 풀이된다.

문항별 응답 시간을 고려한 채점 방식이나, 고려하지 않은 CALT 채점 방식의 타당성을 비교 결과는 다음과 같다. 청해에서는 시간 고려 알고리즘이 비고려 알고리즘보다 TOP와의 상관 관계가 다소 높다. 문법 및 어휘에서는 시간 고려 알고리즘이 비고려 알고리즘보다 TOP와의 상관 관계가 훨씬 더 높다. 이는 문항별 응답 시간을 고려한 채점 방식이 고려하지 않은 CALT 채점 방식보다 전반적 의사소통능력을 더 신뢰성있게 측정함을 보여주는 결과이다. 반면, 독해 영역에서는 시간 고려와 비고려 알고리즘의 TOP와의 상관관계에서 볼 때 큰 차이가 없다. 이는, 다른 영역보다 더 복잡한 인지과정이 요구되는 독해에서는 문항별 응답 시간이 큰 변수로 작용하지 않기 때문으로 해석된다. 이는 시간 변수의 중요성을 고려한 Oller (1995)와 최인철 (2000), Choi (2002)의 연구의 결과와 맥을 같이 한다. 본 연구에서, 언어습득 이론과 수차례의 시뮬레이션 연구를 바탕으로 하여 시간 변수를 적절하게 고려하도록 알고리즘을 개발하였기 때문에 바람직한 결과가 나온 것으로 해석된다.

PBLT과 TOP간의 상관관계 결과는 다음과 같다. PBLT의 청해시험은 TOP의 전반적인 의사소통가능도와 타시험 영역보다 높은 상관 관계를 보이고 있는데, 이는 청해시험의 내용 타당도뿐만 아니라 시험 방식의 특성 (음성언어로만 제시 들려주

는 방식)에 기인한 것으로 풀이된다. 또한, 청해시험과 TOP 시험의 발음간의 상관관계가 PBLT의 타 영역보다 높음을 알 수 있었는데, 이는 외국인 학습자가 습득한 발음과 청해력간의 관계가 높음을 알 수 있다. 이는 청력 장애자의 발음을 고려해 볼 때에도 이해할 수 있는 결과이다. 또한, 살아있는 발음을 많이 청해함으로써 청해력을 키우고 발음도 좋아질 수 있는 외국어 습득 이론과도 일맥상통하는 것이다 (Gilbert, 1959). PBLT의 문법시험 결과는 TOP의 타 영역보다 문법과 높은 상관관계를 보이고 있는데, 이는 살아있는 의사소통능력의 밑바탕이 되는 잠재적인 구문능력 (Acquisition)을 측정하는데 속도화 시험 (speeded test: Oller, 1995)방식이 타당함을 보여주는 결과로 해석된다.

좀더 세부적인 영역간의 관계를 살펴보면 다음과 같다. TOP 발음영역과 PBLT 청해시험, 시간고려 CALT 청해, 그리고 비시간고려 CALT 청해시험 사이의 상관관계, TOP 유창성영역과 나머지 3가지 청해시험 사이의 상관관계를 보면 큰 차이가 없음을 알 수 있다 (0.8161, 08242, 0.8163; 0.8182, 0.7863, 0.7823). 이는 청해시험에서는 PBLT에서도 모든 문제를 음성언어로 제시하는 시험 방식상 수험자 개인별로 수험시간에 차이가 없기 때문에, PBLT와 CALT간의 큰 차이가 없는 것으로 풀이된다. 이로써, PBLT의 청해시험에서 모든 문제를 음성언어로만 제시하는 방식이 수험자의 발음 능력을 간접적으로 측정할 수 있는 타당한 시험방식임을 보여주는 결과로 해석된다.

또한, TOP의 문법영역과 PBLT의 문법시험, 시간 고려 알고리즘인 CALT 문법시험, 그리고 시간 비고려 CALT의 문법시험간의 상관관계를 살펴보면, 전체적으로 시간 제약이 있는 PBLT의 문법시험과 TOP 문법간의 상관관계가 (0.8473)이 다른 두가지 시험과 TOP 문법간의 상관관계 (0.7414, 0.7209)보다 많이 높은 것으로 나타났다. 이는 전반적으로 시간 제약이 있는 지필방식의 문법시험에 잘 적응한 반면, 생소한 방식의 CALT에 잘 적응하지 못한 것에 기인한 것으로 보인다. 수험자들이 CALT 시험을 완료하고 처음 치루는 CALT시험에 대한 느낌 중에 수험자 불안감이 큰 것으로 연구자에게 밝혔는데, 문자언어로 제시된 시험중에 제일 먼저 제시된 문법시험에 적응하는데 다소 어려움을 겪은 데 기인한 것으로 풀이되며, 시간 변수와 멀티미디어 CALT 시험 방식이 수험자에게 미치는 불안감을 포함한 전반적인 수험전략 및 태도 문제에 대한 보다 심도 깊은 연구가 필요하다.

표 2. CALT, PBLT, TOP의 상관관계 (전반부)

Corr	PTOT	PLC	PGR	PVC	PRC	C1TOT	C1LC	C1GR	C1VC	C1RC
PTOT	1	.9564^	.8223^	.8680^	.9485^	.8948^	.7934^	.7864^	.8041^	.7599^
PLC	.9564^	1	.7866^	.8342^	.8235^	.8632^	.7927^	.7512^	.7870^	.7015^
PGR	.8223^	.7866^	1	.7157^	.7173^	.8067^	.6845^	.7199^	.7364^	.7184^
PVC	.8680^	.8342^	.7157^	1	.7795^	.8665^	.7821^	.7179^	.8454^	.7173^
PRC	.9485^	.8235^	.7173^	.7795^	1	.8128^	.6980^	.7265^	.7066^	.7173^
C1TOT	.8948^	.8632^	.8067^	.8665^	.8128^	1	.9143^	.7091^	.8457^	.8767^
C1LC	.7934^	.7927^	.6845^	.7821^	.6980^	.9143^	1	.5591^	.7783^	.6385^
C1GR	.7864^	.7512^	.7199^	.7179^	.7265^	.7091^	.5591^	1	.7535^	.5501^
C1VC	.8041^	.7870^	.7364^	.8454^	.7066^	.8457^	.7783^	.7535^	1	.6311^
C1RC	.7599^	.7015^	.7184^	.7173^	.7173^	.8767^	.6385^	.5501*	.6311^	1
C2TOT	.8889^	.8581^	.8106^	.8620^	.8048^	.9979^	.9101^	.7067^	.8462^	.8774^
C2LC	.7925^	.7904^	.6846^	.7802^	.6985^	.9139^	.9983^	.5546*	.7757^	.6409^
C2GR	.7296^	.6979^	.7090^	.6656^	.6652^	.6627^	.4965*	.9783^	.7160^	.5284*
C2VC	.7513^	.7423^	.7200^	.7988^	.6457^	.8048^	.7265^	.7364^	.9691^	.6061^
C2RC	.7597^	.7012^	.7166^	.7164^	.7176^	.8762^	.6376^	.5505^	.6282^	.9999^
TPR	.8137^	.8161^	.7083^	.7973^	.7121^	.8497^	.8242^	.6376^	.7738^	.6710^
TFL	.8403^	.8182^	.7659^	.7094^	.7709^	.8620^	.7863^	.6805^	.7103^	.7396^
TGR	.8141^	.8257^	.6473^	.7040^	.7299^	.8195^	.7381^	.7414^	.8006^	.6570^
TSL	.7668^	.7561^	.7332^	.7809^	.6663^	.8845^	.7681^	.6021^	.7811^	.8234^
TOC	.8192.	.7769^	.7236^	.7664^	.7635^	.9160^	.8716^	.6346^	.8248^	.7595^

N of case 35 2-tailed Signif: * - .01 ^ - .001

NB:

PTOT	'PBLT TOTAL'
PLC	'PBLT Listening Comprehension'
PGR	'PBLT Grammar'
PVC	'PBLT Vocabulary'
PRC	'PBLT Reading Comprehension'
C1TOT	'CALT TOTAL TIMED'
C1LC	'CALT Listening Comprehension'TIMED'
C1GR	'CALT Grammar TIMED'
C1VC	'CALT Vocabulary TIMED'
C1RC	'CALT Reading Comprehension TIMED'
C2TOT	'CALT TOTAL UNTIMED'
C2LC	'CALT Listening Comprehension UNTIMED'
C2GR	'CALT Grammar UNTIMED'
C2VC	'CALT Vocabulary UNTIMED'
C2RC	'CALT Reading Comprehension UNTIMED'
TPR	'TOP Pronunciation'
TFL	'TOP Fluency'
TGR	'TOP Grammmar'
TSL	'TOP Sociolinguistic Competence'
TOC	'TOP Overall Comprehensibility'.

표 3. CALT, PBLT, TOP의 상관관계 (후반부)

Corr	C2TOT	C2LC	C2GR	C2VC	C2RC	TPR	TFL	TGR	TSL	TOC
PTOT	8889^	.7925^	.7296^	.7513^	.7597^	.8137^	.8403^	.8141^	.7668^	.8192^
PLC	8581^	.7904^	.6979^	.7423^	.7012^	.8161^	.8182^	.8257^	.7561^	.8069^
PGR	8106^	.6846^	.7090^	.7200^	.7166^	.7083^	.7659^	.8473^	.7332^	.7536^
PVC	8620^	.7802^	.6656^	.7988^	.7164^	.7973^	.7094^	.7040^	.7809^	.7464^
PRC	8048^	.6985^	.6652^	.6457^	.7176^	.7121^	.7709^	.7299^	.6663^	.7335^
C1TOT	9979^	.9139^	.6627^	.8048^	.8762^	.8497^	.8620^	.8195^	.8845^	.9160^
C1LC	9101^	.9983^	.4965*	.7265^	.6376^	.8242^	.7863^	.7381^	.7681^	.8716^
C1GR	7067^	.5546*	.9783^	.7364^	.5505^	.6376^	.6805^	.7414^	.6021^	.8346^
C1VC	8462^	.7757^	.7160^	.9691^	.6282^	.7738^	.7103^	.6006^	.7811^	.8248^
C1RC	8774^	.6409^	.5284*	.6061^	.9999^	.6710^	.7396^	.6570^	.8234^	.7595^
C2TOT		.9123^	.6711^	.8192^	.8769^	.8447^	.8617^	.8187^	.8951^	.9100^
C2LC	9123^	1	.4956*	.7298^	.6401^	.8163^	.7823^	.7339^	.7753^	.8628^
C2GR	6711^	.4956*1		.7431^	.5289^	.5926^	.6627^	.7209^	.5972^	.7881^
C2VC	8192^	.7298^	.7431^	1	.6037^	.7359^	.6934^	.5853^	.7814^	.7787^
C2RC	8769^	.6401^	.5289*	.6037^	1	.6696^	.7307^	.6582^	.8217^	.7576^
TPR	8447^	.8163^	.5926^	.7359^	.6696^	1	.7861^	.7456^	.7209^	.8467^
TFL	8617^	.7823^	.6627^	.6934^	.7407^	.7861^	1	.8265^	.8253^	.8524^
TGR	8187^	.7339^	.7209^	.7853^	.6582^	.7456^	.8265^	1	.7306^	.8233^
TSL	8951^	.7753^	.5972^	.7814^	.8217^	.7209^	.8253^	.7306^	1	.8034^
TOC	9100^	.8628^	.5881^	.7787^	.7576^	.8467^	.8524^	.8233^	.8034^	1

3.2. 문맥 효과

수험자 마다 상이한 문항이 상이한 순서로 제시될 때에, 수험결과에 미치는 효과를 고려하는 것이 필요하다. 즉, 시험 영역별, 소재별, 유형 (과제)별로 상이한 문항를 제시하는 순서가 수험 결과에 영향을 미칠 수 있다. 이런 소위 문맥 효과를 극소화 하기 위해서, 본 연구에서는 영역별, 유형별 제시 순서를 지필고사의 것과 유사하게 하여, 수험자들에 게 미치는 문맥 영향을 균등화하였다. 문맥 효과를 고려한 내용 균형화에 대해서 좀더 체계적인 차후 연구가 필요하다.

또한, 수험자의 실력수준별로 상이한 문항 순서 효과도 함께 고려하는 것이 필요하다. 이는, 중간 난이도의 다소 쉬운 문항부터 어려운 문항이 제시될 고급 실력의 수험자에게는 큰 문제가 되지 않겠지만, 어려운 문항부터 쉬운 문항의 순서로 제시될 경우에는 하급 실력의 수험자에게는 큰 영향을 미칠 것으로 예상된다. 따라서, 본 연구에서는 중간 난이도의 문항들을 먼저 제시하고, 초기 4-5문항 내에 수

험자의 능력을 추정한 후, 바로 수험자의 능력에 적합한 난이도의 문항을 제시하였다.

3.3. 신뢰도

컴퓨터의 개별 적응시험은 측정의 정확성 즉 신뢰도를 전제하고 있는 IRT에 근거한 시험 방식이고, 측정 오차의 최소화를 알고리즘에서 중요시한다. 본 연구에서는 모든 피험자의 경우에 측정오차가 .1이하로 나타났는데, 측정오차는 신뢰성와 반비례하는 관계이므로, 이로써 본 연구에서 개발한 알고리즘이 적절한 문항수과 함께 높은 수준의 신뢰도를 보이고 있음을 알 수 있다.

또한, 신뢰도 및 객관도에 영향을 미칠 수 있는 요인으로 고려해야 할 것이 시험 영역별 제시 순서이다. 즉, 수험자 개인별 취향을 고려해서 영역별 문항 제시 순서를 수험자가 마음대로 정할 수 있게 하는 것이 좋다는 주장이 있지만, 이는 신뢰도에 미치는 영향을 고려해 볼 때에 바람직하지 않다. 최인철 (2000)연구 결과와 피험자들에 대한 면담을 통해서 시험 마지막에 독해 시험을 볼 때에 피로 효과 때문에 적지 않은 어려움을 겪게 되었음을 알 수 있었다. 이렇듯이 영역별 문항 제시 순서는 수험 결과에 영향을 미칠 수 있다. 따라서, 수험 결과를 동일한 여건에서 시험을 치르게 함으로써 가능한 한 신뢰도를 극대화하기 위해서 영역별 제시 순서를 동일하게 제시하는 것이 바람직하다.

4. 결 론

CALT 알고리즘개발을 위한 타당한 이론적 배경을 근거로 한 주요 고려사항 (초기 문항의 난이도, 능력 모수 추정 방법, 문항 추출 알고리즘, 시험 길이, 측정 오차, 문맥 효과, 문항별 응답 시간 등)에 대한 체계적인 분석을 통하여 시뮬레이션 및 실제 피험자 자료에 근거한 연구를 통하여 보다 바람직한 알고리즘을 개발할 수 있었다. 기존의 교육 평가 연구의 대부분은 시뮬레이션을 근거로 하여 알고리즘을 개발하므로, 실제 시험 시행에서 지필고사보다 낮은 신뢰도을 얻는 경우가 많으므로, 반드시, 현실적으로 여러 제약이 있기는 하지만 실제 자료에 근거한 연구를 통하여 보다 타당한 결과를 얻을 수 있음을 알 수 있었다. 특히, 알고리즘 개발을 위한 모든 고려사항에서 이론적으로 타당한 방법들의 장점을 통합하는 절충적인 방법이 실제 자료를 근거로 한 실험 연구를 통해서 보다 바람직함을 알 수 있었다.

본 연구를 통하여 개발된 CALT 알고리즘을 활용할 때에, 대부분의 수험자들에게는 컴퓨터 개별 적응 시험이 지필시험 방식보다 전반적인 의사소통능력을 보다 타당하게 측정하고 있음을 알 수 있었다. 또한, 문항별 응답 시간을 고려한 채점 방식이 고려하지 않은 CALT 채점 방식보다 전반적 의사소통능력을 더 신뢰성있게 측정함을 알 수 있었다. 이로써, 지필고사와 달리 문항별 시간 변수를 측정할 수 있는 CALT의 특성을 고려하는 것이 바람직하다는 것을 알 수 있었다. 또한, 지필고

사 방식과 TOP 시험에서 측정하는 의사소통능력 요소간의 상관관계 결과는, 문법의 속도화 시험방식과 순수하게 음성언어로 제시되는 청해시험 방식의 타당성을 입증하는 중요한 자료이다. 특히, TOP 발음과 유창성 영역과 PBLT 청해시험, CALT 청해 사이의 상관관계는 TEPS의 청해시험에서 모든 문제를 음성언어로만 제시하는 방식이 수험자의 발음 능력을 간접적으로 측정할 수 있는 타당한 시험방식임을 보여주는 결과로 해석된다.

끝으로, 화면과 음성을 통해 제시되는 CALT 시험은 PBLT와 시험방식양상이 매우 크게 상이하므로, 이런 멀티미디어적 변수가 수험 결과에 미치는 영향에 대한 체계적인 연구와 더나아가서 멀티미디어 기술과 전산 언어학 (음성인식과 자연언어처리 등) 기술을 접목하여 표현능력 (말하기, 쓰기)을 측정하는 수행평가 도구의 개발이 차후 연구 과제로 중요하게 대두될 전망이다.

참 고 문 헌

백순근 · 채선희 · 홍미영 · 임재훈 · 전은화. (1998). 컴퓨터를 이용한 학력검사 개발 연구. 한국교육과정평가원. 연구보고 RRE 98-4.

최인철. (1998). Test of Oral Proficiency의 개발. 어학연구, 34(1), 245-286.

최인철. (2000). 영어 의사소통능력의 모의 구술 면접 (Simulated Oral Proficiency Interview) 시험 방식 양상 타당성 검증. 응용언어학 16(1), 215-246.

Bachman, L. F., Davidson, F., Ryan, K, and I-C. Choi. (1995). *Studies in Language Testing 1: An Investigation into the Comparability of Two Tests of English as a Foreign Language.* Cambridge: Cambridge University Press.

Brennan R. L. (1992). *Elements of Generalizability Theory.* Iowa City: The American College Testing Program.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology* 1(1), 44-59.

Choi, I-C. (1992). *Theoretical Studies in Second Language Acquisition: An Application of Item Response Theory to Language Testing.* New York: Peter Lang Publishing Inc.

Choi, I-C. (1994). Content and construct validation of a criterion-referenced English proficiency test. *English Teaching* 48, 311-348.

Choi, I-C. (1995). A Comparability Study on SNUCREPT and TOEIC. *Language Research* 31(2), 357-386.

Choi, I-C. (1997). Essential test method facets of a general English proficiency test and their validity as perceived by test-takers. *Language Research* 33(4), 773-799.

Choi, I-C. (2002). Significance of response time in computer-based/adaptive language tests 5(1), 9-31.

Choi, I-C. and L. F. Bachman. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing, 9*(2).

Choi, I-C, Kim, K. S., and J. Y. Boo. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing* 20(3), 294-319.

Dunkel, P. (1991). The effectiveness research on computer-assisted instruction and computer assisted language learning. In P. Dunkel, Ed., *Computer-assisted language learning and testing: Research issues and practice* (pp. 5-36). New York: Newbury House.

Gilbert, J. (1995) Pronunciation practices as an aid to listening comprehension. In D. J. Mendelson and J. Rubin, Eds., *A Guide for the Teaching of Second Language Listening* (pp. 97-112). San Diego: Dominic Press.

Hambleton, R. K., and H. Swaminathan (1984). *Item Response Theory: Principles and Applications.* Hingham, MA: Kluwer, Nijhoff.

Hambleton, R. K., Swaminathan, H., and H. J. Rogers. (1991). *Fundamentals of Item Response Theory.* Thousand Oaks, CA: Sage.

Hansen, D. N. (1969). An investigation of computer-based science testing. In R. C. Atkinson and H. A. Wilson Eds., *Computer-assisted Instruction: A Book of Readings.* (pp. 209-226). New York: Academic Press.

Hulin, C. L., Drasgow, F., and C. K. Parsons. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood, IL: Dow-Jones, Irwin.

Kingsbury, G. G. and A. R. Zara. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-75.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J. M. (1999). *A User's Guide to FACETS.* Chicago,IL: MESA Press.

Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement* 8, 147-151.

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika 36,* 227-241.

Lord, F. M. (1977). A broad range tailored test of verbal ability. *Applied Psychological Measurement,* 95-100.

Mazzeo, J. and A. L. Harvey. (1988). *The Equivalence of Scores from Auto-*

mated and Conventional Educational and Psychological Tests: A Review of the Literature. Report No. CBR 87-8, ETS RR 88-21. Princeton, NJ: Educational Testing Service.

Mead, A. D. and F. Drasgow. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin 114*, 449-458.

Oller, J. W. Jr. (1995). Review of Content and Construct Validation of a Criterion-referenced English Proficiency Test by Choi (1994). *English Teaching 50*(3), 161-168.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association* 70, 351-356.

Russel, M. and W. Haney. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives* 5(3).

Sands, W. A., Waters, B. K., and J. R. McBride (1997). *Computerized Adaptive Testing: from Inquiry to Operation.* Washington D.C.: American Psychological Association.

Stout, William, R. Nandakumar, B. Junker, and H. H. Chang. (1991). *Dimtest and Testsim.* Urbana, IL: University of Illinois.

Thissen, D. and R. J. Mislevy. (1990). Testing algorithms. In H. Wainer, Ed., *Computerized adaptive testing: a Primer,* (pp. 103-135). NJ: Lawrence Erlbaum Associations, Publishers.

Thissen, D., Steinberg, L., and A. R. Fitzpartrick. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement* 26, 161-176.

Wainer, H. and G. L. Kiely. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement* 24, 185-201.

Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, and D. Thissen. (Eds). (1990). *Computerized Adaptive Testing: A Primer.* Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J. Steinberg, L., and D. Thissen (Eds). (2000). *Computerized Adaptive Testing: A Primer.* 2nd Edition. Hillsdale, NJ: Lawrence Erlbaum.

Wetzel, C. D., and J. R. McBride. (1983). *The Influence of Falliable Item Parameters on Test Information during Adaptive Testing* (TR 83-15). San Diego, CA: Navy Personnel Research and Development Center.

Weiss, D. J. (1974). *Strategies of Adaptive Ability Measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Psychometric Methods Program.

Weiss, D. J. (Ed). (1983). *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing.* New York: Academic Press.

Weiss, D. J. and G. G. Kingsbury. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement* 21, 361-375.

<부록> CALT 알고리즘 능력 모수 추정 결과 예

1. DB 기록 (각 문항별 능력 모수 추정 결과 및 응답 시간 등)

1 번
문제번호 : 2167 / A
Theta : M0250
변환값 : -0.25

Value : 0.722
Avalue : 0.562
Bvalue : -0.211
남은문제수 : Sx(-1)Tx(-1) << 난이도(theta) : -0.25 (M0250) 2 2 1 10
결과=> 정오 : 0 Theta : -0.25 / NextP :-0.25 시간:7

---

2 번
문제번호 : 1391 / A
Theta : M0750
변환값 : -0.75

Value : 0.722
Avalue : 0.593
Bvalue : -0.774
남은문제수 : Sx(-2)Tx(-1) << 난이도(theta) : -0.75 (M0750) 2 2 1 4
결과=> 정오 : 1 Theta : -0.75 / NextP :-0.75 시간:1

---

3 번
문제번호 : 355 / B
Theta : M0625
변환값 : -0.625

Value : 0.722
Avalue : 0.593
Bvalue : -0.576
남은문제수 : Sx(-3)Tx(-2) << 난이도(theta) : -0.625 (M0625) 2 2 1 10
결과=> 정오 : 1 Theta : -0.625 / NextP :-0.625 시간:1

---

4 번
문제번호 : 1048 / D
Theta : M0625
변환값 : -0.625

Value : 0.722
Avalue : 0.606

Bvalue : -0.598
남은문제수 : Sx(-4)Tx(-1) << 난이도(theta) : -0.625 (M0625) 2 2 1 9
결과=> 정오 : 1 Theta : -0.5 / NextP :-0.5 시간:1

---

5 번
문제번호 : 762 / A
Theta : M0500
변환값 : -0.5

Value : 0.722
Avalue : 0.593
Bvalue : -0.448
남은문제수 : Sx(-5)Tx(-1) << 난이도(theta) : -0.5 (M0500) 2 2 1 3
결과=> 정오 : 0 Theta : 0.185 / NextP :0.125 시간:1

---

6 번
문제번호 : 1042 / C
Theta : P0125
변환값 : 0.125

Value : 0.722
Avalue : 0.568
Bvalue : 0.145
남은문제수 : Sx(-6)Tx(-1) << 난이도(theta) : 0.125 (P0125) 2 2 1 1
결과=> 정오 : 1 Theta : 0.389 / NextP :0.375 시간:1

---

7 번
문제번호 : 1397 / A
Theta : P0375
변환값 : 0.375

Value : 0.722
Avalue : 0.593
Bvalue : 0.363
남은문제수 : Sx(-7)Tx(-1) << 난이도(theta) : 0.375 (P0375) 2 2 1 11
결과=> 정오 : 1 Theta : 0.853 / NextP :0.875 시간:1

---

8 번
문제번호 : 75 / A
Theta : P0875

변환값 : 0.875
Value : 0.722
Avalue : 0.506

Bvalue : 0.912
남은문제수 : Sx(-8)Tx(-1) << 난이도(theta) : 0.875 (P0875) 2 2 1 12
결과=> 정오 : 0 Theta : 0.626 / NextP :0.625 시간:1

---

9 번
문제번호 : 6566 / D
Theta : P0625
변환값 : 0.625

Value : 0.722
Avalue : 0.337
Bvalue : 0.554
남은문제수 : Sx(-1)Tx(-3) << 난이도(theta) : 0.625 (P0625) 3 2 1 10
결과=> 정오 : 1 Theta : 0.582 / NextP :0.625 시간:1

---

10 번
문제번호 : 6946 / A
Theta : P0625
변환값 : 0.625

Value : 0.722
Avalue : 0.241
Bvalue : 0.739
남은문제수 : Sx(-2)Tx(-2) << 난이도(theta) : 0.625 (P0625) 3 2 1 4
결과=> 정오 : 1 Theta : 0.82 / NextP :0.875 시간:1

중략

---

2. 능력 모수 추정 과정

Theta -0.25

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
2167	0.562	-0.211	M0250	0	0	0	0		

Delta Theta 0/-0.25

Theta -0.75

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
1391	0.593	-0.774	M0750	1	0	0	0		

Delta Theta 0/-0.75

Theta -0.625

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
355	0.593	-0.576	M0625	1	0	0	0		

Delta Theta 0/-0.625

Theta -0.5

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
1048	0.606	-0.598	M0625	1	0	0	0		

Delta Theta 0/-0.5

Theta -0.5

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
2167	0.562	-0.211	M0250	0	0.431	-0.243	0.077	0.722	1.176
1391	0.593	-0.774	M0750	1	0.568	0.255	0.086	1.444	0.832
355	0.593	-0.576	M0625	1	0.519	0.285	0.087	2.166	0.679
1048	0.606	-0.598	M0625	1	0.525	0.287	0.091	2.888	0.588
762	0.593	-0.448	M0500	0	0.486	-0.289	0.087	3.61	0.526

Delta Theta 0.406/-0.094

Theta -0.094

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
2167	0.562	-0.211	M0250	0	0.527	-0.297	0.078	4.332	0.48
1391	0.593	-0.774	M0750	1	0.664	0.198	0.078	5.054	0.444
355	0.593	-0.576	M0625	1	0.619	0.225	0.082	5.776	0.416
1048	0.606	-0.598	M0625	1	0.626	0.226	0.085	6.497	0.392
762	0.593	-0.448	M0500	0	0.588	-0.349	0.085	7.219	0.372

Delta Theta 0.211/0.117

Theta 0.117

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
2167	0.562	-0.211	M0250	0	0.577	-0.325	0.077	7.941	0.354
1391	0.593	-0.774	M0750	1	0.71	0.171	0.072	8.663	0.339
355	0.593	-0.576	M0625	1	0.667	0.196	0.078	9.385	0.326
1048	0.606	-0.598	M0625	1	0.676	0.196	0.08	10.107	0.314
762	0.593	-0.448	M0500	0	0.638	-0.379	0.081	10.829	0.303

Delta Theta 0.078/0.195

중략

Theta 0.054

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
616	0.425	0.205	P0250	0	0.472	-0.201	0.045	29.6	0.183
1095	0.452	-0.907	M0875	1	0.676	0.146	0.044	30.322	0.181
1566	0.456	-0.816	M0750	1	0.662	0.153	0.046	31.044	0.179
1567	0.565	-0.511	M0500	1	0.632	0.207	0.074	31.766	0.177
4970	0.543	-0.255	M0250	1	0.57	0.233	0.072	32.488	0.175
1559	0.559	0.49	P0500	0	0.397	-0.223	0.074	33.21	0.173
1730	0.456	0.689	P0750	1	0.379	0.283	0.048	33.932	0.171
364	0.328	1.172	P1250	0	0.349	-0.115	0.024	34.654	0.169
5751	0.425	1.025	P1000	0	0.331	-0.141	0.04	35.376	0.168
700	0.532	0.755	P0750	0	0.346	-0.185	0.064	36.098	0.166
685	0.465	0.501	P0500	0	0.412	-0.192	0.052	36.82	0.164
2083	0.328	0.34	P0250	0	0.46	-0.151	0.026	37.542	0.163
3859	0.569	-0.005	M0000	0	0.514	-0.293	0.08	38.264	0.161
1619	0.569	-0.174	M0125	0	0.554	-0.316	0.079	38.986	0.16

683	0.484	-0.866	M0875	1	0.68	0.154	0.05	39.708	0.158
3329	0.55	-0.625	M0625	1	0.653	0.19	0.068	40.43	0.157
323	0.445	-0.37	M0375	0	0.579	-0.258	0.048	41.152	0.155
2943	0.465	-0.338	M0375	1	0.576	0.196	0.052	41.874	0.154
13504	0.448	-0.075	M0125	1	0.524	0.213	0.05	42.596	0.153
6316	0.425	0.093	M0000	1	0.492	0.215	0.045	43.317	0.151

Delta Theta 0.014/0.068

Theta 0.068

문항	AV	BV	theta	정오	PV	분자	분모	IV	추정오차
616	0.425	0.205	P0250	0	0.475	-0.202	0.045	44.039	0.15
1095	0.452	-0.907	M0875	1	0.679	0.145	0.044	44.761	0.149
1566	0.456	-0.816	M0750	1	0.664	0.152	0.046	45.483	0.148
1567	0.565	-0.511	M0500	1	0.635	0.205	0.073	46.205	0.147
4970	0.543	-0.255	M0250	1	0.573	0.231	0.072	46.927	0.145
1559	0.559	0.49	P0500	0	0.401	-0.225	0.075	47.649	0.144
1730	0.456	0.689	P0750	1	0.381	0.281	0.049	48.371	0.143
364	0.328	1.172	P1250	0	0.35	-0.116	0.024	49.093	0.142
5751	0.425	1.025	P1000	0	0.333	-0.142	0.04	49.815	0.141
700	0.532	0.755	P0750	0	0.349	-0.186	0.064	50.537	0.14
685	0.465	0.501	P0500	0	0.415	-0.194	0.052	51.259	0.139
2083	0.328	0.34	P0250	0	0.462	-0.152	0.026	51.981	0.138
3859	0.569	-0.005	M0000	0	0.517	-0.295	0.08	52.703	0.137
1619	0.569	-0.174	M0125	0	0.558	-0.318	0.079	53.425	0.136
683	0.484	-0.866	M0875	1	0.683	0.153	0.05	54.147	0.135
3329	0.55	-0.625	M0625	1	0.656	0.188	0.068	54.869	0.135
323	0.445	-0.37	M0375	0	0.582	-0.26	0.048	55.591	0.134
2943	0.465	-0.338	M0375	1	0.579	0.195	0.052	56.313	0.133
13504	0.448	-0.075	M0125	1	0.527	0.211	0.05	57.035	0.132
6316	0.425	0.093	M0000	1	0.495	0.214	0.045	57.756	0.131

Delta Theta -0.003/0.065

최 인 철
136-701
서울 성북동 안암동 5가
고려대학교 사범대학 영어교육과
전자우편 : icchoi@korea.ac.kr

접수일자 : 2003. 8. 25
수정본 접수 : 2004. 2. 16
게재결정 : 2004. 2. 21