언어평가에 대한 컴퓨터 기술의 활용방안*

이 영 식
(한남대학교)

Lee, Young Shik. 2003. **A Status Quo Study of Using Computer Technology for Language Testing.** *Korean Journal of English Language and Linguistics* 3-4, 571-588. The purpose of this study is to investigate into the various ways that the computer technology is used for language testing. Three uses of computer technology are mentioned: 1) computer-adaptive language testing and computer-based language testing, 2) the scoring of performance-based language assessment, and 3) the development and use of psychometric tools for analyzing the scoring results. Although the various uses of computer technology could provide expanded possibilities for language testing development, the developers should be reminded that they are currently subject to indepth research which could support their validity. In this regard, the advantages and limitations of some uses of computer technology for language testing are discussed.

Key Words: language assessment, computer technology, PhonePass computer-based testing (CBT), computer-adaptive testing (CAT), e-rater, psychometrics, item response theory (IRT), G-theory

1. 서론

전세계적으로 의사소통능력능력 평가에 대해 컴퓨터를 활용하여 실시하고 있는 방법도 현대의 언어평가의 경향이다. 특히 컴퓨터에 의한 TOEFL시험이 이미 우리나라에서도 최근에 실시되고 있어서, 이제는 우리도 컴퓨터를 통한 언어평가가 현실임을 알 수 있다. 따라서 우리는 언어평가에 대한 컴퓨터의 활용에 대해 연구 조사함으로써 우리의 언어평가 실시 및 관리에 대한 안목을 제고할

*본 논문은 2002년도 한남대학교 학술연구 조성비 지원에 의하여 연구되었음

필요가 있다. 이에 따라 우리는 언어평가에 대해 컴퓨터를 활용하는 실제 사례에 대해 우선 조사하여 그 유형을 분류하여 파악하여 볼 필요가 있다. 그러한 조사를 바탕으로 언어평가의 기저 이론과 컴퓨터에 대한 구체적인 운영체제, 활용과정 및 방법에 대해 기술 분석할 뿐 만 아니라, 언어평가에 대한 컴퓨터의 활용에 대한 이해를 도모하여 우리의 언어평가에 대한 컴퓨터 활용에 대한 전문성을 고양할 필요가 있다. 또한 컴퓨터를 활용한 언어평가가 만능은 아니어서 각각의 장점과 함께 컴퓨터 언어평가 도구의 한계성도 파악할 필요가 있다. 결국 그러한 장점과 한계성을 각각 비교 조사함으로써 효율적인 언어평가를 위한 컴퓨터 활용이 우리 실정에 정착될 수 있는 안목을 기를 수 있다고 본다.

2. 언어평가에 대한 컴퓨터 기술의 활용

21세기를 맞이하여 모든 인간의 생활이 정보화가 가속화됨에 따라, 언어교육 및 언어평가도 정보 및 컴퓨터 기술(information and computer technology)을 활용하기 시작하여, 급기야는 컴퓨터를 활용한 언어교육과 평가에 대한 연구와 논문이 많이 나오게 되었다. 필자가 보건대, 컴퓨터를 언어평가에 이용하는 방법은 대략적으로 다음과 같이 세 가지로 분류하여 볼 수 있다: 1) 컴퓨터를 시험방법으로 이용하는 방법; 2) 컴퓨터를 채점 판단에 동원하는 방법; 3) 시험 채점결과에 대한 분석을 위하여 통계 및 심리측정 프로그램으로 컴퓨터를 활용하는 방법. 따라서 본 논문에서는 위와 같은 컴퓨터 활용방법을 논하고, 그 의의를 살펴보고자 한다.

2.1. 컴퓨터 적응시험(computer-adaptive testing: CAT) 및 컴퓨터 기반시험(computer-based testing: CBT)

언어평가에 대한 컴퓨터의 활용은 주로 객관식 평가를 위한 컴퓨터 적응시험(computer-adaptive testing: CAT)과 컴퓨터 기반 시험(computer-based testing: CBT)을 들 수 있다. 지난 세기동안 컴퓨터의 많은 발달로 개인 컴퓨터가 많이 보급되고, 또한

컴퓨터를 이용한 평가가 많은 연구와 발전을 거듭하여 이제는 실용화가 되었다. 특히 미국 ETS(Educational Testing Service)에 의한 TOEFL(Test of English as a Foreign Language) 시험이 지필고사(paper-and-pencil test) 방식으로 치루었다가, 지난 10년간 많은 연구와 실험을 거친 후 전세계적으로 컴퓨터를 이용한 시험(ETS, 1998; Wang 등, 2001)으로 대체하였으며, 우리나라도 2000년 가을부터 컴퓨터를 이용한 시험으로 치루게 되었다.

Dunkel(1997, p. 8)은 컴퓨터 적응 시험이란 컴퓨터의 발달된 기술을 이용하여 수험자의 능력에 시험문항의 난이도를 맞추는 시험방법으로, 근본적으로 맞춤시험(tailored testing)이라고 정의하였다. 따라서 컴퓨터 적응시험이란 시험문항의 난이도에 수험자의 능력을 맞추도록 고도의 컴퓨터 기술이 동원된 평가방법을 말하며, 근본적으로 시험문항을 수험자의 능력에 맞추어서 실시하는 맞춤식 평가라 할 수 있다. 특히 평가이론의 중요한 발전이라 할 수 있는 문항반응이론(item response theory)은 컴퓨터 적응시험을 가능하게 하였다. 이러한 컴퓨터 적응 시험을 언어평가에 이용하게 되었으며, 이러한 언어평가를 컴퓨터 적응 언어시험(computer-adaptive language test: CALT)이라고 부른다.

컴퓨터 적응 시험은 우선 문제은행(item bank)의 구축을 필요로 한다. 이러한 문제은행을 구축을 위하여 각 문항들이 단일능력 난이도 척도(a single ability-difficulty scale)에 의거 조정(calibrate)되고, 이러한 각각의 문항의 난이도에 따라 문항들의 범위가 정해지고 결국 시험을 치루는 수험자들이 그에 따라 배치되게 된다는 원리이다. 이러한 문제은행의 원리는 전술한 바와 같이 문항반응이론에 바탕을 두고 있으며, 각 문항의 난이도는 사전에 시험을 치루어 – 대부분의 경우에는 지필고사 방식으로 치루어진 것으로 – 결과가 도출되고, 이러한 각 문항들의 난이도는 미리 프로그램화되어 컴퓨터 적응 시험에 입력이 된다. 이러한 적응적 알고리듬(adaptive algorithm)은 각각의 수험자가 주어진 문항에 대해 응답하는 양상에 따라 문제은행으로부터 가장 적절한 수준의 난이도 문항을 선택하여 주는 절차를 지니고 있다.

수험자가 시험을 시작할 때 첫 문항은 중간정도 난이도 수준을 지니고 있으나, 수험자가 첫 문항을 해결하면 다음에는 그 보다

약간 어려운 문항이 나오고, 그 다음 두번째 문항을 해결하게 되면 더욱 어려운 문항이 나오고, 해결하지 못하면 그보다 쉬운 문항이 나오게 된다. 반면에 수험자가 첫 문항을 해결하지 못하게 되면 약간 쉬운 문항이 나오게 되며, 또 두 번째 문항을 해결하지 못하게 되면 더욱 쉬운 문항이 나오게 되나, 그 문항을 해결하게 되면 그보다는 약간 어려운 문항이 나온다. 즉, 문항선별 알고리듬(item selection algorithm)에 의해 처음에 무작위적으로 나오는 문항을 각각의 수험자가 해결하게 되면 더욱 어려운 문항이 나오고, 해결하지 못하면 더욱 쉬운 문항이 나오는 과정을 여러번 거치다보면, 결국 수험자 수준에 맞는 문항만 나오게 되어 그 수험자의 능력이 평가된다는 것이다.

이렇게 시험문항이 수험자의 능력에 따라 다르게 등장하여 각 수험자가 치르는 시험문항이 다른 수험자와 치르는 문항과 다를 수 있으나, 한편 다수의 수험자가 여러 번 시험을 치르는 과정에서 같은 문항을 치를 수 있다. 비록 시험의 알고리듬이 무작위적으로 문항이 등장하도록 고려하였지만, 우연히 같은 문항이 어느 한 집단에게 자주 나오게 되면 시험문제의 누출이 염려될 소지가 있다. 이러한 시험문제의 보안을 위하여 고안된 알고리듬이 있게 된다. Chalhoub-Deville과 Deville(1999, p. 287)에 의하면 노출 통제 모수(exposure control parameters)가 컴퓨터 적응시험에 이용되며, 이는 시험에 자주 나오게 되는 문항들을 통제하여 과도하게 특정 시험문항이 편중되어 나오지 않게 하는 장치라고 설명하였다. 이는 수험자에게 주로 많이 등장하는 문항을 통제하여 문제은행에 있는 문항들이 골고루 사용되게 하거나 또는 시험의 보안을 유지하기 위한 목적으로 특정 문항의 유출을 방지하기 위한 것이다.

한편 우리는 여기에서 CAT(computer-adaptive test)와 CBT (computer-based test)에 대한 용어의 정의를 분명히 할 필요가 있다. 이 두 용어의 주요한 차이는 CAT는 문제선별과 같은 적응적 알고리듬을 지니고 있으나, CBT는 반드시 그렇지 않고 단순히 컴퓨터를 이용하여 치루는 시험을 의미할 수 있다. CBT는 지필고사와 같은 내용과 방법을 단지 컴퓨터를 통하여 치르는 시험으로 CAT를 포괄하는 개념이라 할 수 있고, 결국 CAT는 CBT의 한 방법이라 할 수 있다. 그리하여 CAT는 우리말로 '컴퓨터

적응시험'이라고 번역할 수 있으며, CBT는 '컴퓨터 이용시험' 또는 '컴퓨터 기반시험'이라고 번역할 수 있다. CAT에서의 적응 알고리듬은 한번에 단지 한 문항만 제시되고, 컴퓨터가 다음 문항을 고르기 전에 그 문항을 채점하기 때문에 수험자는 그 문항에 대한 답을 해결하지 않고 그냥 넘어가지 못한다. 또한 수험자가 답을 확인하고 일단 기입을 한 후에는 그 문항으로 되돌아 올 수 없게 된다. 그러나 CBT에서는 적응 알고리듬을 지니지 않고 선형적(linear)으로 문항이 제시되므로 수험자가 문항을 그냥 넘어갈 수 있고, 나중에 다시 그 문항으로 되돌아 올 수도 있다. 현재 컴퓨터를 이용한 TOEFL시험은 CBT라고 부르며, 이러한 시험의 세 분야 중에서 듣기(Listening)와 문법(Structure)은 적응 방식이고, 독해(Reading)는 적응 방식이 아닌 선형 방식이다.

1) 장점

컴퓨터 적응시험은 각 문항들이 문제은행으로부터 컴퓨터 알고리듬에 의해 개별 수험자에 맞는 문제만 나오게 되어있어서 맞춤시험(tailored test)라는 장점을 지니고 있다. 따라서 Bergstrom과 Gershon (1994, p. 25)은 다음과 같이 말하였다

> When the difficulty of items is targeted to the ability of candidates, maximum information is obtained from each item for each candidate, so test length can be shortened without loss of reliability.

시험문항의 난이도가 수험자의 능력에 맞추어지므로, 지필고사와 같이 처음부터 끝까지 주어진 아주 쉽거나 어려운 문제를 불필요하게 수험자가 치룰 필요가 없어서, 수험자의 시험시간을 상당히 절약할 수 있다. 이것은 결국 짧은 시간에 평가할 수 있으면서도 수험자 능력에 대한 최대 정보가 각각의 수험자를 위해 주어질 수 있어서, 수험자에게 중복된 수준의 문항을 계속 풀어야 하는 시험의 잉여성(redundancy)을 방지할 수 있다. 또한 수험자가 자기 능력보다 너무 어려운 문제나 쉬운 문제를 계속 풀어야 할 필요가 없기 때문에 시험을 통하여 수험자에게 심리적으로

좌절감이나 무력감을 심어주는 것을 방지할 수 있다. 아울러 컴퓨터 적응시험의 알고리듬은 각각의 수험자에게 다른 문항을 제공하므로 지필고사처럼 옆 수험자의 문제를 보면서 답을 맞출 수 있는 소지를 미리 없애주어서 시험의 부정행위를 방지할 수 있다. 특히 시험을 치룬 후에 즉각적인 결과(성적)가 나온다는 것이 커다란 장점으로, 수험자의 능력에 대한 진단평가로서의 좋은 평가도구가 될 수 있다.

또한 컴퓨터 기반 시험은 일정한 시간에 일정한 장소에 모여 치룰 필요가 없는 장점이 있다. 특히 컴퓨터 기반 시험은 대부분 인터넷을 통하여 시험을 치룰 수 있어서 웹기반 시험(web-based test)이 가능하도록 하였고, 수험자가 언제 어디에서든지 등록하여 즉각 시험을 치룰 수 있으며 또한 인터넷을 통하여 그 결과를 즉각 통보 받을 수 있다.

2) 단점

컴퓨터 적응시험은 아직도 다지선다형 방식의 선별된 답에 의존하기 때문에 수험자의 수행능력 평가(performance-based testing)를 하지 못하고 분리시험(discrete testing) 방식에 국한되었다. 즉 컴퓨터 적응시험은 작문(essay)과 인터뷰와 같은 언어수행능력을 포함하지 못함으로써 의사소통 기술(communicative skills)보다는 언어지식(linguistic knowledge)에 국한되는 평가도구에 그치고 말았다. 또한 컴퓨터 적응시험에서의 적응 알고리듬으로 인하여 수험자에게 한번에 오로지 한 문항만 제시되고, 수험자는 그 문항을 해결하지 않고 그냥 넘어갈 수 없으며, 일단 답을 기입한 후에는 그 문항으로 되돌아 올 수 없게 된다. 따라서 수험자가 순간적인 실수로 인하여 오답을 기입하게 되면 정정할 기회가 없게 된다는 단점이 있다.

한편 컴퓨터를 이용한 언어시험이 지필고사와 같은 종래의 언어시험과는 다른 능력을 평가할 수 있는 소지가 있다는 의문도 제기되었고, 급기야는 이러한 컴퓨터를 활용한 언어시험의 타당도에 대해서도 비판이 나오게 되었다(Fulcher, 1999). 컴퓨터를 이용한 시험의 타당도에 대해서 재연구와 조명이 있어야 한다는 주장도 나왔다. 그리하여 Bachman(2000, p. 9)은 다음과 같이

말하였다.

> [T]he new task formats and modes of presentation that multi-media computer-based test administration makes possible raise all the familiar validity questions, and may require us to redefine the very constructs we believe we are assessing.

또한 수험자가 컴퓨터나 인터넷을 통하여 시험을 치루게 되면, 시험결과가 수험자 본인의 언어능력뿐만 아니라 수험자가 이용하는 컴퓨터의 성능에 따라 크게 달라질 소지가 많다. 즉 어떤 수험자가 컴퓨터를 이용한 시험에서 성적이 좋은 결과가 나왔을 때, 그 수험자의 의사소통능력이 우수해서 좋은 결과가 나올 수도 있지만, 뿐만 아니라 수험자의 컴퓨터 기술이나 수험자가 치루는 컴퓨터의 성능이 좋아서 다른 수험자와 다른 시험결과가 나올 수 있는 소지가 있다. 따라서 컴퓨터 기반시험은 선별시험(screening test)보다는 수험자 자신평가(self-assessment)나 진단평가(diagnostic test)에 적합하고, 고부하 시험(high-stakes)보다는 저부하(low-stakes) 또는 중부하(medium-stakes) 시험에 적합할 수 있다. 따라서 Roever (2001, p. 84)는 웹기반 시험에 대해 다음과 말하고 있다.

> It is argued that WBTs are most appropriate in low-stakes testing situations; but with proper supervision, they can also be used in medium-stakes situations although they are not generally recommended for high-stakes situations.

컴퓨터를 이용한 시험이 과거 수십년 동안 연구되고 발전하여 이제는 실용화가 되었지만, 그러나 아직도 많은 한계가 있다는 것이 지적되었다. 그리하여 Chalhoub-Deville과 Deville(1999, p. 292)은 다음과 같이 말하였다.

> Although CBT has developed significantly over the past

decade it is not a testing panacea and must be viewed with its current limitations (particularly expense and technological complexity) in mind.

따라서 컴퓨터를 이용한 언어시험이 수험자 언어능력 평가에 대한 모든 것을 해결해 주지는 못하는 것도 현실로써 인정할 필요가 있다. 아울러 이러한 시험에 부과되는 많은 비용과 컴퓨터 기술의 복잡성도 결코 간과하여서는 안된다.

2.2. 채점 판단에 대한 컴퓨터의 이용

1) 컴퓨터 음성인식 기술을 이용한 영어 말하기 평가(PhonePass)

지금까지 컴퓨터가 시험절차에 이용되는 방법에 대해 논의하였는데, 컴퓨터 기반시험이나 컴퓨터 적응시험은 주로 영어독해나 문법 및 어휘와 같이 분리적으로 측정되는 수용기술(receptive skills)에만 관심을 두었다. 그러한 연유에는 기술적 제약으로 인하여 컴퓨터를 이용한 말하기 평가가 어렵게 여겨졌기 때문이다. 그러나 이제는 컴퓨터가 수용기술에 대한 시험절차로서의 평가뿐만 아니라, 수행평가로서 수험자가 치른 시험결과에 대한 채점도 담당하고 있는 컴퓨터 기술이 많이 제시되고, 그에 따른 많은 개발연구와 실험결과가 나오고 있다. 이미 다지선다형 객관식 시험의 결과를 컴퓨터가 자동적으로 채점하고 있다는 것은 오래전 주지의 사실이지만, 수행평가와 같은 영어 말하기나 작문시험에서도 이미 컴퓨터의 자동채점이 실시되고 있으며 그에 대한 연구결과가 괄목하다고 볼 수 있다.

특히 영어 말하기 평가에서의 컴퓨터의 활용은 음성인식 기술을 수반하고 있으며, 그러한 음성인식 기술은 영어 말하기의 여러 양상들을 자동적으로 평가하여 채점할 수 있게 하였다. 예를 들면 PhonePass (Ordinate Corporation, 1998; Hubbard, 1999)는 영어 말하기에 대한 가능한 모든 요인들, 즉 주로 미국영어(American English)에 대한 듣기, 유창성, 발음, 문법, 어휘 등을 측정하도록 동원된 음성인식기술에 바탕을 두고 있다(Bernstein, 1997; 2000). 이러한 명칭이 말하는 바와 같이, PhonePass는 컴퓨터 시스템을 이용한 전화를 통하여 영어 말하기 평가가 실시되며, 1) 소리내어

읽기(reading aloud), 2) 문장 반복(repeating sentences), 3) 반대말 대기(naming opposite words), 4) 짧은 답하기(providing short answers), 5) 개방형 응답하기(giving open responses)와 같은 다섯 가지의 유형으로 구성되어 있다.

그러나 이러한 영어 말하기 평가체제도 고급수준의 영어 말하기 능력을 평가하지 못하고, 단지 단순하고 기계적인 대화 양상만을 측정하고 있는 실정이다. 특히 영어 말하기의 비예측성이나 대화의 협상에 대한 고려가 전혀 없어서 과연 진정한 영어말하기 평가도구라고 보기에는 아직도 어렵다고 말할 수 있다.

한편 영어 말하기와 같은 복잡한 수행평가를 현실적으로 실시할 때－특히 외국어 상황에서의 영어 말하기 평가를 실시할 때－인간채점(human rating)에 대한 신뢰도의 문제가 대두되는 바, 이에 대한 채점의 비신뢰성을 보완하여 줄 수 있다고 PhonePass의 제작자는 주장하고 있다(Bernstein, 2000). 특히 그에 의하면, PhonePass에 의한 영어 말하기 평가가 비록 단순하고 수준이 높지 못하더라도, 평가결과는 수험자의 전반적 영어 말하기 능력에 대한 상당한 상관관계를 지니고 있다고 한다. 또한 영어평가가 평가자체만으로 그치지 않고, 그 평가의 내용과 과정이 영어교육의 내용과 방법에 대해 많은 파급효과(washback)를 미치는 바, 영어말하기 평가의 전문성과 채점의 신뢰성 부족으로 인하여 우리 영어교육이 기본수준부터 아예 말하기 시험을 기피하는 것이 우리 현실이라면, 이러한 영어 말하기 컴퓨터 평가기술을 활용하여 우리의 영어 말하기 평가 방법에 적용하여 보는 것도 좋은 시도라고 볼 수 있다. 여하튼 이와 같은 컴퓨터 기술의 발전이 언어평가의 발전에 미치는 잠재력은 상당히 고무적이고, 향후 연구와 개발노력의 추이가 기대된다고 하겠다.

2) 영어작문에 대한 컴퓨터 채점(e-rater)

영어 채점에 대한 컴퓨터 기술의 활용으로 영어작문에 대한 컴퓨터 채점의 가능성이 열렸고, 그 대표적인 예로서 현재 ETS 고유의 채점 시스템이라 할 수 있는 e-rater를 들 수 있다. Burstein 등(1998)은 GMAT의 13개 영어작문 문제에 대해 e-rater의 채점과 인간 전문가들의 채점을 비교한 결과 87-94%의 일치성

(agreement)을 보였다는 연구결과를 내놓았다. 또한 Test of Written English (TWE)의 두개 작문과제에서도 e-rater의 채점이 인간 전문가 채점과 93-94%의 일치를 보였다고 주장하였다. 이러한 통계는 2명의 전문 채점자 사이의 채점결과와 거의 일치하는 것으로, e-rater에 의한 채점이 고부하(high-stakes) 시험에서도 제2채점자(second reader)로서 채점신뢰도 뿐만 아니라, 영어작문 채점에 소요되는 상당한 인력, 시간과 경비를 절감할 수 있다고 그들은 주장하였다.

TWE에 대한 e-rater채점은 인간 채점처럼 전체적 채점방식(holistic scoring)에 바탕을 두고 있으며, 채점 기준은 통사적인 면(syntactic structure and syntactic variety), 담화적인 면(discourse cues and organization of ideas), 그리고 주제분석과 어휘사용(topical analysis and vocabulary usage)에 근거하고 있으며, 최고점은 6점으로 최하점은 1점으로 구성되어 있다. 그 후 이러한 TWE의 영어작문 채점에 대해서 한편 두명의 인간 전문 채점자의 사이의 상관관계가 .75이나, 이러한 e-rater와 각각의 인간 전문 채점자 사이의 채점 상관관계가 .73이며, 인간채점에 의한 점수와 e-rater에 의한 등급점수가 92% 일치한다는 연구결과(Burstein & Chodorow, 1999)도 나왔다. 특히 그들의 연구에 의하면 75%의 영어작문이 비원어민(non-native speakers)에 의해 쓰여 졌는데, 그러한 영어작문의 채점에 이용된 채점특성이 원어민(native speakers)에 쓰여진 영어작문 채점에 이용된 채점특성과 차이가 없음을 주장하였다. e-rater에 의한 채점이 원어민에 의한 영어작문 뿐만 아니라 비원어민에 의한 영어작문에도 잘 적용되며, 또한 e-rater가 비표준 영어 통사구조나 담화구조에 의해서도 혼동(confound)되지 않는다고 주장하였다.

이상에서 보건대, 작문에 대한 e-rater와 같은 컴퓨터 채점의 가능성이 현실화되어 영어작문평가에 대한 새로운 가능성을 열어주었다고 볼 수 있다. 그러나 여기에서 우리는 e-rater에 대해 유의하여야 할 것은, 컴퓨터 채점이 인간 채점자를 대체하기에는 아직도 현실적으로 많은 문제가 있을 수 있다. 대규모의 표준화 고부하 시험(standardized high-stakes exam)에서는 인간 채점자를 많이 동원하여 복수로 채점하는 것이 관례이며, 그러한 대량의

인간 복수채점이 반드시 높은 신뢰도를 유지하는 것은 아니면서 시간과 경비면에서 효율성이 떨어지는 것도 사실이다. 그리하여 컴퓨터 채점을 통하여 인간 채점의 여러 측면을 보완할 수 있다는 주장이 앞에서 언급한 것처럼 제기되어 왔지만, 그것은 다만 실험연구의 주장에 불과할 수 있고 아직도 컴퓨터 채점의 타당도에 많은 문제가 완전히 해결된 상태라고 볼 수 없다. 특히 영어작문 채점의 중요한 요인인 주제내용(topical content)에 대한 담화차원의 분석에 대해서는 컴퓨터 채점의 타당성이 아직도 문제가 될 수 있어서, 컴퓨터 채점은 시간적으로 경제적으로 인간 채점을 보완하는 제2의 채점자에 불과할 수밖에 없다. 그러나 한편 향후 영어작문 채점에 대한 발전을 위하여 앞으로도 인간채점과 컴퓨터 채점을 사용하여 상호 비교하는 연구를 계속함으로써 영어작문 채점에 대한 타당성을 제고할 뿐만 아니라, 시험실시의 실제적인 측면(비용과 시간)을 보완할 수 있는 방안을 도모하고, 더 나아가서 학습자의 작문진단과 교육적인 피드백을 자동 제공하는 모형구성에도 중요한 역할을 할 것으로 기대한다.

한편 미국 UCLA의 ESLPE (English as a Second Language Placement Examination) 시험도 컴퓨터 채점을 실시하면서 현재 연구 중에 있는데, 특기할 만한 것은 컴퓨터를 이용한 시험에서의 개방형 질문에 대해서도 자동적 컴퓨터 채점도 실시 가능하다는 것이다. 이 시험 채점은 개방형 질문에 대한 모든 가능한 답안을 채점 전에 나열하여 컴퓨터가 수험자의 답안과 주어진 정답들과 짝지어 맞추어서 실시한다(matching). 물론 이러한 컴퓨터 채점의 단점으로서 사전에 입력된 답안만 채점하는 한계를 들 수 있지만, 그러한 답안을 채점하는 과정이 주어진 답안은 정확히 맞추기 때문에 인간채점과는 달리 상당한 채점의 정확성을 지니면서도 많은 시간과 경비를 절감할 수 있다는 것이다. 특히 개방 응답형 문항을 채점을 할 때에 정답을 미리 정해놓고 채점을 시작하므로 채점과정에서 새로운 정답이 불가피하게 나오게 되어 이미 실시한 채점에 대해 수정하는 재채점을 실시할 필요가 있는데, 컴퓨터에 의한 재채점은 더욱 신속하고 인력을 절감할 수 있다는 효과가 있다(Carr and Xi, 2002).

2.3. 측정프로그램에 대한 컴퓨터 활용

언어평가에 대한 컴퓨터의 활용은 많은 통계 및 측정프로그램들이 컴퓨터 프로그램으로 개발되어 평가에 대한 신뢰도나 타당도를 검증하는데 사용되는 것이다. 대부분의 측정프로그램들이 고도의 연산과정을 수반하기 때문에 종래에는 사람의 수작업에 의거한 계산으로 어렵거나 거의 불가능하다고 여겼으나, 최근에 컴퓨터 기술이 발달하고 또한 컴퓨터의 사용이 급증함에 따라 복잡한 측정도구들이 컴퓨터 프로그램으로 만들어져서 많이 활용하게 되었다(Bachman, 1991; 2000).

특히 지난 세기동안 모든 것들은 측정될 수 있고 수량화할 수 있다는 심리측정학자들에 의한 주장에 많은 기초를 두고 측정이론에 대한 많은 연구를 한 결과 이제는 고도의 측정프로그램들이 많이 출현하게 되었으며, 그 대표적인 예로 문항반응이론(item response theory: IRT)에 기초를 둔 프로그램과 일반화 가능도 이론(generalizability theory: g-theory)에 바탕을 둔 프로그램을 들 수 있다.

문항반응이론은 크게 두가지 주류가 있는데, 우선 첫번째 주류는 Lord와 Novick(1968)의 고전적인 교과서 *Statistical Theories of Mental Test Scores*로 거슬러 올라갈 수 있다. 이 책은 Allen Birnbaum이 저술한 문항반응이론의 네 단원을 포함하고 있으며, 종래의 고전검사이론(classical test theory: CTT)과 달리 심리측정분야에서의 획기적인 업적으로 인정되기도 한다. 그 후 Darrell Bock이 문항반응이론 모형의 모수를 효율적으로 추정하는 알고리듬의 개발에 관심을 가졌다(Bock, 1972). 특히 Bock은 모수들을 추정할 수 있는 주변 최대 우도 추정법(marginal maximum likelihood method)을 개발하였고(Bock and Aitken, 1981), 그러한 문항반응이론의 역사에 대한 논문을 쓰기도 하였다(Bock, 1997). 그에 따라 문항반응이론의 많은 전문가들이 컴퓨터 프로그램을 개발하였는데, 그 예로 BILOG, TESTFACT, XCALIBRE, MULTILOG, PARSCALE, RUMM 등을 들 수 있다(Embretson & Reise, 2000). 이러한 문항반응이론 컴퓨터 프로그램은 앞에서 언급한 바와 같이 컴퓨터 적응시험의 개발을 가능하게 하였다.

문항반응이론의 또 한 주류는 덴마크의 수학자 Georg Rasch에 의해 개발되고 미국 시카고 대학의 Ben Wright에 의해 널리 보급된 일모수 모형(one-parameter model)으로 Rasch모형을 들 수 있다. 원래 기본적 라쉬모형은 다지선다형 시험과 같은 이분문항(dichotomous items)의 분석에 국한되었으나, 그 후에 많은 연구가 이루어져 채점등급(rating scales)과 같은 데이터의 분석에도 확장되었다. 특히 Linacre(1989; 1994)에 의한 다국면적 라쉬 측정(many-facet Rasch measurement)에서는 채점자와 같은 평가의 제반 국면에 대한 분석을 가능하게 하였으며, 이러한 이론에 의거 Facets라는 컴퓨터 프로그램(Linacre and Wright, 1990; 1992; 1993)이 개발되었다. 이러한 프로그램은 채점자들에 의해 산출된 점수에 대해 채점자 특성과 시험문항 특성이 미치는 영향을 정교하게 분석할 수 있는 계기를 마련 해주어서 현재 평가 전문가들에 의해 수행평가의 결과 분석에 많이 이용되고 있다(McNamara, 1996; 이 영식, 1998). 이상에서 보는 바와 같이 첫 번째 주류의 문항반응이론 컴퓨터 프로그램은 주로 다지선다형(multiple-choice) 문항 분석에 이용되고 있으면서, 다른 주류의 문항반응이론에 의한 Facets라는 프로그램은 수행평가의 결과 분석에 이용되고 있는 데에 괄목할 만하다.

일반화 가능도 이론(generalizability theory: g-theory)은 고전검사 이론의 연장이라고 볼 수 있으나, 측정오류(measurement error)를 보는 시야에서 많은 차이가 있다. 고전검사이론은 오류를 단지 단일실체(single entity)로 간주하는 반면에, 일반화 가능도 이론은 여러 종류의 오류를 개별적으로 분석하고, 그러한 개별적 오류가 ANOVA 통계적 절차를 이용하여 전반적인 오류로 기여하는 정도를 추정하는 것이다. 그리하여 일반화가능도 이론의 창시자라 할 수 있는 Robert Brennan (2001, p. 4)은 다음과 같이 말한다.

> Perhaps the most important aspect and unique feature of generalizability theory is its conceptual framework. Among the concepts are *universe of admissible observations* and G *(generalizability)* studies, as well as *universes of generalization* and D *(decision)* studies.

이러한 모형은 시험에서 문항이나 과제(task)의 수 뿐만 아니라 수행평가에서의 채점자의 수를 조정하여 야기되는 효과를 추정하는데 유용한 도구로 밝혀졌다. 이러한 이론에 의거 개발된 컴퓨터 프로그램으로 GENOVA를 들 수 있다.

Bachman(1991; 2000)과 Kunnan(1999)은 지난 20세기말 언어평가와 관련된 발전의 한 모습을 컴퓨터를 통한 더욱 세련되고 정교한 측정 및 평가도구의 개발 및 활용이라고 언급하였다. 아무튼 다양한 통계 및 측정이론이 컴퓨터의 발달과 함께 많은 컴퓨터 통계 및 측정 소프트웨어를 개발하였고, 그러한 컴퓨터 소프트웨어 프로그램이 언어평가결과에 대한 분석에 사용되어 타당도 검증방법으로 이용되고 있는 것이 세계적인 추세라고 하겠다.

3. 결론

지금까지 우리는 언어평가에 대한 컴퓨터 기술을 활용한 방안을 조사하여 보았다. 현재 국내에서의 일반인들의 실제 생활에 컴퓨터의 활용은 이미 상당한 수준에 이르렀으나, 언어평가에 대한 컴퓨터의 안목과 활용은 아직 미미한 상태라고 볼 수 있다. 아직도 우리 수험생들에게는 컴퓨터를 활용한 언어평가가 두려움의 대상이 되고, 우리의 언어교육 현장에서는 컴퓨터를 활용한 언어평가 방식을 제대로 활용할 수 있는 안목이 부족하다고 볼 수 있다. 따라서 컴퓨터를 활용한 언어평가 방식에 대한 조사 연구는 이제까지 지필고사 위주로 실시된 객관식 위주의 언어평가가 수험자의 언어구사력을 제한된 방법으로 평가할 수밖에 없는 우리 현실에 보다 다양한 언어평가 방식을 도입할 수 있는 안목과 실질적 계기를 마련할 수 있다고 본다.

컴퓨터를 통한 언어평가는 단순한 문장이나 글 위주의 문항에서 벗어나 다양한 그림을 동원하고 음성위주의 시험문항을 개발할 수 있다. 또한 수험자가 시험을 치르는 동안 각 문항의 난이도가 수험자에게 조정되어 수험자 수준보다 너무나 어렵거나 쉬운 문항을 치를 필요가 없으며, 개별적으로 수험자가 서로 다른

시험문항을 동시에 다발적으로 치를 수 있는 환경을 구축할 수 있는 바탕을 마련할 수 있다. 이러한 컴퓨터를 활용한 언어평가는 인터넷으로도 가능하여 결국 수험자 스스로 공부하고 평가할 수 있는 교육환경을 조성할 수 있다.

컴퓨터를 활용한 언어평가가 영어 말하기와 작문과 같은 수행평가에도 확대될 수 있어서, 이러한 언어평가에 대한 컴퓨터 활용의 연구는 언어평가의 새로운 가능성과 다양성을 이해할 수 있는 바탕을 마련하여 줄 수 있다. 아울러 수행평가의 채점은 채점자의 주관성으로 인하여 수행평가의 채점신뢰도가 큰 문제가 될 수 있으나, 컴퓨터를 통한 다양한 측정도구의 활용으로 채점의 주관성과 오류를 상당히 검증할 수 있고, 수행평가의 신뢰성과 채점의 객관성을 유지할 수 있는 바탕을 마련하여 줄 수 있다. 이러한 언어평가방법과 검증은 시험을 치르는 학생들의 언어구사력을 바르게 평가할 수 있는 바탕을 마련하여 주어서 결국 의사소통을 위한 언어평가 뿐만 아니라 언어교육 내용이나 과정에 바람직한 결과를 유도할 수 있는 계기도 마련하여 줄 수 있다.

그러나 컴퓨터를 활용한 언어평가나 언어교육이 반드시 만능은 아니어서, 언어평가에 대한 컴퓨터 활용 연구는 그러한 한계를 제대로 인식할 수 있는 안목을 바르게 제공할 필요가 있다. 따라서 우리는 언어평가에 대한 컴퓨터 운영체제와 소프트웨어를 바르게 활용할 수 있는 언어평가 및 교육 관계자들의 이해를 향상시킬 수 있도록 우리는 항상 노력할 필요가 있다. 특히 현재 국내에서는 언어평가에 대한 컴퓨터 이용 개발과 연구를 비판적으로 추진하여, 향후 언어평가에 대한 컴퓨터의 오용과 남용을 잘 조절할 수 있는 능력을 갖추는 것도 더욱 중요하다고 본다. 결국 그러한 이해를 통하여 우리 언어교육과 평가의 다양성과 타당성을 재고하고 고양시킬 수 있는 계기를 마련하여 줄 수 있다고 기대한다.

참고문헌

이영식. 1998. 영어작문평가의 채점신뢰도에 대한 분석. ≪영어교육≫ 53, 179-200.

Bachman, L. F. 1991. What does language testing have to offer? *TESOL*

Quarterly 25, 671-704.

Bachman, L. F. 2000. Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17, 1-42.

Bachman, L. F. and A. Cohen. 1998. *Interface Between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press.

Bergstrom, B. and R. Gershon. 1994. Computerized adaptive test for licensure and certification. *CLEAR Exam Review*, 25-27.

Bernstein, J. 1997. Computer-based oral proficiency assessment: Field test results. Paper presented at the Language Testing Research Colloquium. Orlando, Florida, March 1997.

Bernstein, J. 2000. Fully automatic, semi-automatic, and fully human spoken language tests. Paper presented at Applied Linguistics Association of Korea 2000 Summer International Conference "Applied Linguistics: New Millennium, New Paradigm", International Studies Hall, Korea University, Seoul, Korea, June, 23-24.

Bock, R. D. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 29-51.

Bock, R. D. 1997. A brief history of item response theory. *Educational Measurement: Issues and Practices* 16, 21-33

Bock, R. D. and Aitken. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46, 443-459.

Brennan, R. L. 2001. *Generalizability Theory.* New York, NY: Springer.

Burstein, J. and M. Chodorow. 1999. Automated essay scoring for nonnative English speakers. Princeton, NJ: Educational Testing Service.

Burstein, J., K. Kukich, S. Wolff, C. Lu, and M. Chodorow. 1998. *Computer Analysis of Essays.* Princeton, NJ: Educational Testing Service.

Carr, N. T. and X. Xi. 2002. Construct refinement and automated scoring in web-based testing. Paper presented at 24th Language Testing Research Colloquium, Hong Kong Polytechnic University, 12th-15th December.

Chalhoub-Deville, M. and C. Deville. 1999. Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics* 19, 273-299.

Chalhoub-Deville, M., ed., 1999. *Issues in Computer Adaptive Testing of Reading Proficiency.* New York: Cambridge University Press.

Chapelle, C. 2001. *Computer Applications in Second Language Acquisition.* Cambridge University Press.

Cumming, A. and R. Berwick., eds., 1996. *Validation in Language Testing.*

Modern Languages in Practice 2. Clevedon Avon: Multilingual Matters.

Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley, and T. McNamara. 1999. *Dictionary of Language Testing: Studies in Language Testing 7.* Cambridge: Cambridge University Press.

Douglas, D. and C. Chapelle, eds., 1993. A new decade of language testing research. Selected papers from the 1990 language testing research Colloquium. Alexandria, VA: TESOL.

Dunkel, P. A., ed., 1991. *Computer-assisted Language Learning and Testing.* New York: Newbury House.

Dunkel, P. A. 1997. Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher JALT* 21, 7-13.

Educational Testing Service. 1998. *TOEFL 1998 Products and Services Catalogue.* Princeton, NJ: ETS.

Embretson, S. E. and S. P. Reise. 2000. *Item Response Theory for Psychologists.* Malwah, NJ: Lawrence Erlbaum Associates.

Fulcher, G. 1999. Computerizing a language placement test. *ELT Journal* 53, 289-299.

Hanson-Smith, E. 2001. Computer-assisted language learning. In R. Carter and D. Nunan, eds., *The Cambridge Guide to Teaching English to Speakers of Other Languages.* Cambridge University Press.

Hubbard, P. 1999. A review of Ordinate's PhonePass. *ESL Magazine Product Review.*

Kunnan, A. J. 1999. Recent developments in language testing. *Annual Review of Applied Linguistics* 19, 235-253

Linacre, J. M. 1989. *Many-faceted Rasch Measurement.* Chicago, IL: MESA Press.

Linacre, J. M. 1994. An introduction to many-facet Rasch measurement. Paper presented at the pre-colloquium workshop on FACETS, Language Testing Research Colloquium '94, Washington DC, 4th March.

Linacre, J. M. and B. D. Wright. 1990. *FACETS.* Chicago, IL: MESA Press.

Linacre, J. M. and B. D. Wright. 1992. *Facets: Rasch Measurement Computer Program,* Version 2.6. Chicago, IL: MESA Press.

Linacre, J. M. and B. D. Wright. 1993. *A User's Guide to FACETS,* Version 2.6. Chicago, IL: MESA Press.

Lord, F. N. and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

McNamara, T. F. 1990. Item response theory and the validation of an ESP test for health professionals. *Language Testing* 7, 52-75.

McNamara, T. F. 1996. *Measuring Second Language Performance.* London: Longman.

McNamara, T. F. 2000. *Language Testing.* Oxford: Oxford University

Press.
Ordinate Corporation. 1998. *PhonePass Test Validation Report*. Menlo Park, CA: Ordinate.
Roever, C. 2001. Web-based language testing. *Language Learning & Technology* 5, 84-94.
Sands, W. A., B. K. Waters, and McBride, eds., 1997. *Computerized Adaptive Testing: From Inquiry to Operation*. Washington, DC: American Psychological Association.
Taylor, C., J. J. Jamieson, and D. Eignor. 1997. Measuring the effects of computer familiarity on computer-based language tasks. Paper presented at the Language Testing Research Colloquium, Orlando, Florida.
Taylor, C., J. J. Jamieson, D. Eignor, and I. Kirsch. 1998. The relationship between computer familiarity and performance on computer-based TOEFL test tasks: TOEFL Research Report 61. Princeton, NJ: ETS.
Wang, X. B., D. Eignor, M. Golub-Smith, Y. Lee, and P. Carey. 2001. Computer-based TOEFL: A discussion of some technological and psychometric issues. Paper presented at 23rd Annual Language Testing Research Colloquium. Marriott Pavilion Hotel, St. Louis, Missouri. Feb 20-24, 2001.

이영식
대전광역시 대덕구 오정동 133번지
한남대학교 영어교육과
우편번호: 306-791
전화번호: 041) 629-7415
E-mail: yshlee@mail.hannam.ac.kr

접수일자: 2003. 9. 10.
게재결정: 2003. 11. 26.